TURING 图灵新知

思考的乐趣

Matrix67数学笔记

顾森　著



人民邮电出版社
POSTS & TELECOM PRESS

数字版权声明

图灵社区的电子书没有采用专有客户端，您可以在任意设备上，用自己喜欢的浏览器和PDF阅读器进行阅读。

但您购买的电子书仅供您个人使用，未经授权，不得进行传播。

我们愿意相信读者具有这样的良知和觉悟，与我们共同保护知识产权。

如果购买者有侵权行为，我们可能对该用户实施包括但不限于关闭该帐号等维权措施，并可能追究法律责任。



顾森

网名Matrix67，北京大学中文系应用语言学专业学生，数学爱好者。2005年开办数学博客http://www.matrix67.com，至今已积累上千篇文章，已有上万人订阅。长期为各类科普杂志供稿，从事中学数学教育工作多年。

图灵新知

思考的乐趣

Matrix67数学笔记

顾森 著

人民邮电出版社
北京

图书在版编目（ＣＩＰ）数据

思考的乐趣 : Matrix67数学笔记 / 顾森著. -- 北京 : 人民邮电出版社, 2012.6（2017.3重印）
（图灵新知）
ISBN 978-7-115-27586-8

Ⅰ. ①思… Ⅱ. ①顾… Ⅲ. ①数学－普及读物 Ⅳ. ①01-49

中国版本图书馆CIP数据核字(2012)第043369号

内 容 提 要

本书内容大多是从作者6年多以来积累的上千篇博客中节选而来的，分为“生活中的数学”、“数学之美”、“几何的大厦”、“精妙的证明”和“思维的尺度”五部分。书中基本不涉及高深的数学理论，但是内容新颖、时尚，既有与现实生活联系紧密的应用型话题，又有打通几何、代数联系的富有启发性的讨论，还间或介绍了一些著名数学难题的最新研究进展，信息十分丰富。

本书是广大数学爱好者的美味佳肴，只要具备简单数学基础即能阅读。

◆ 著　　　　顾　森
责任编辑　明永玲
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京　　　　　　印刷
◆ 开本：700×1000　1/16
印张：17.5
字数：286千字　　　　2012年6月第1版
印数：58 001－59 000册　　　　2017年3月北京第17次印刷

定价：45.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第8052号

序一

我本不想写这个序。因为知道多数人看书不爱看序言。特别是像本书这样有趣的书，看了目录就被吊起了胃口，性急的读者肯定会直奔那最吸引眼球的章节，哪还有耐心看你的序言?

话虽如此，我还是答应了作者，同意写这个序。一个中文系的青年学生如此喜欢数学，居然写起数学科普来，而且写得如此投入又如此精彩，使我无法拒绝。

书从日常生活说起，一开始就讲概率论教你如何说谎。接下来谈到失物、物价、健康、公平、密码还有中文分词，原来这么多问题都与数学有关！但有关的数学内容，理解起来好像并不是很容易。一个消费税的问题，又是图表曲线，又是均衡价格，立刻有了高深模样。说到最后，道理很浅显：向消费者收税，消费意愿减少，商人的利润也就减少；向商人收税，成本上涨，消费者也就要多出钱。数学就是这样，无论什么都能插进去说说，而且千方百计把事情说个明白，力求返璞归真。

如果你对生活中这些事无所谓，就从第二部分开始看吧。这里有“让你立刻爱上数学的 8 个算术游戏”。作者口气好大，区区 5 页文字，能让人立刻爱上数学? 你看下去，就知道作者没有骗你。这些算术游戏做起来十分简单却又有趣，背后的奥秘又好像深不可测。8 个游戏中有 6 个与数的十进制有关，这给了你思考的空间和当一回数学家的机会。不妨想想做做，换成二进制或八进制，这些游戏又会如何? 如果这几个游戏勾起了探究数字奥秘的兴趣，那就接着往下看，后面是一大串折磨人的长期没有解决的数学之谜。问题说起来很浅显明白，学过算术就懂，可就是难以回答。到底有多难，谁也不知道。也许明天就有人想到了一个巧妙的解答，这个人可能就是你；也许一万年仍然是个悬案。

但是这一部分的主题不是数学之难，而是数学之美。这是数学文化中常说常新的话题，大家从各自不同的角度欣赏数学之美。陈省身出资两万设计出版了《数学之美》挂历，十二幅画中有一张是分形，是唯一在本书这一部分中出现的主题。这应了作者的说法："讲数学之美，分形图形是不可不讲的。"喜爱分形图的读者不妨到网上搜索一下，在图片库里有丰富的彩色分形图。一边读着本书，一边欣赏神秘而惊人美丽的艺术作品，从理性和感性两方面享受思考和观察的乐趣吧。此外，书里还有不常见的信息，例如三角形居然有 5000 多颗心，我是第一次知道。看了这一部分，马上到网上看有关的网站，确实是开了眼界。

作者接下来介绍几何。几何内容太丰富了，作者着重讲了几何作图。从经典的尺规作图、有趣的单规作图，到疯狂的生锈圆规作图、意外有效的火柴棒作图，再到功能特强的折纸作图和现代化机械化的连杆作图，在几何世界里我们做了一次心旷神怡的旅游。原来小时候玩过的折纸剪纸，都能够登上数学的大雅之堂了！最近看到《数学文化》月刊上有篇文章，说折纸技术可以用来解决有关太阳能飞船、轮胎、血管支架等工业设计中的许多实际问题，真是不可思议。

学习数学的过程中，会体验到三种感觉。

一种是思想解放的感觉。从小学里学习加减乘除开始，就不断地突破清规戒律。两个整数相除可能除不尽，引进分数就除尽了；两个数相减可能不够减，引进负数就能够相减了；负数不能开平方，引进虚数就开出来了。很多现象是不确定的，引进概率就有规律了。浏览本书过程中，心底常常升起数学无禁区的感觉。说谎问题，定价问题，语文句子分析问题，都可以成为数学问题；摆火柴棒，折纸，剪拼，皆可成为严谨的学术。好像在数学里没有什么问题不能讨论，在世界上没有什么事情不能提炼出数学。

一种是智慧和力量增长的感觉。小学里使人焦头烂额的四则应用题，一旦学会方程，做起来轻松愉快，摧枯拉朽地就解决了。曾经使许多饱学之士百思不解的曲线切线或面积计算问题，一旦学了微积分，即使让普通人做起来也是小菜一碟。有时仅仅读一个小时甚至十几分钟，就能感受到自己智慧和力量的增长。十几分钟之前还是一



头雾水，十几分钟之后豁然开朗。读本书的第四部分时，这种智慧和力量增长的感觉特别明显。作者把精心选择的巧妙的数学证明，一个接一个地抛出来，让读者反复体验智慧和力量增长的感觉。这里有小题目也有大题目，不管是大题还是小题，解法常能令人拍案叫绝。在解答一个小问题之前作者说："看了这个证明后，你一定会觉得自己笨死了。"能感到自己之前笨，当然是因为智慧增长了！

一种是心灵震撼的感觉。小时候读到棋盘格上放大米的数学故事，就感到震撼，原来 $2^{64}-1$ 是这样大的数！在细细阅读本书第五部分时，读者可能一次一次地被数学思维的深远宏伟所震撼。一个看似简单的数字染色问题，推理中运用的数字远远超过佛经里的"恒河沙数"，以至于数字仅仅是数字而无实际意义！接下去，数学家考虑的"所有的命题"和"所有的算法"就不再是有穷个对象。而对于无穷多的对象，数学家依然从容地处理之，该是什么就是什么。自然数已经是无穷多了，有没有更大的无穷？开始总会觉得有理数更多。但错了，数学的推理很快证明，密密麻麻的有理数不过和自然数一样多。有理数都是整系数一次方程的根，也许加上整系数 2 次方程的根，整系数 3 次方程的根等等，也就是所谓代数数就会比自然数多了吧？这里有大量的无理数呢！结果又错了。代数数看似声势浩大，仍不过和自然数一样多。这时会想所有的无穷都一样多吧，但又错了。简单而巧妙的数学推理得到很多人至今不肯接受的结论：实数比自然数多！这是伟大的德国数学家康托的代表性成果。

说这个结论很多人至今不肯接受是有事实根据的。科学出版社去年出了一本书名为《统一无穷理论》，该书作者主张无穷只有一个，不赞成实数比自然数多，希望建立新的关于无穷的理论。他的努力受到一些研究数理哲学的学者的支持，可惜目前还不能自圆其说。我不知道有哪位数学家支持"统一无穷理论"，但反对"实数比自然数多"的数学家历史上是有过的。康托的老师克罗内克激烈地反对康托的理论，以致康托得了终身不愈的精神病。另一位大数学家布劳威尔发展了构造性数学，这种数学中不承认无穷集合，只承认可构造的数学对象。只承认构造性的证明而不承认排中律，也就不承认反证法。而康托证明"实数比自然数多"用的就是反证法。尽管绝大多数的数学家不肯放弃无穷集合概念，也不肯放弃排中律，但布劳威尔的构造性数学也被承认是一个数学分支，并在计算机科学中发挥重要作用。

平心而论，在现实世界确实没有无穷。既没有无穷大也没有无穷小。无穷大和无穷小都是人们智慧的创造物。有了无穷的概念，数学家能够更方便地解决或描述仅仅涉及有穷的问题。数学能够思考无穷，而且能够得出一系列令人信服的结论，这是人类精神的胜利。但是，对无穷的思考、描述和推理，归根结底只能通过语言和文字符号来进行。也就是说，我们关于无穷的思考，归根结底是有穷个符号排列组合所表达出来的规律。这样看，构造数学即使不承认无穷，也仍然能够研究有关无穷的文字符号，也就能够研究有关无穷的理论。因为有关无穷的理论表达为文字符号之后，也就成为有穷的可构造的对象了。

话说远了，回到本书。本书一大特色，是力图把道理说明白。作者总是用自己的语言来阐述数学结论产生的来龙去脉，在关键之处还不忘给出饱含激情的特别提醒。数学的美与数学的严谨是分不开的。数学的真趣在于思考。不少数学科普，甚至国外有些大家的作品，说到较为复杂深刻的数学成果，常常不肯花力气讲清楚其中的道理，可能认为讲了读者也不会看，是费力不讨好。本书讲了不少相当深刻的数学工作，其推理过程有时曲折迂回，作者总是不畏艰难，一板一眼地力图说清楚，认真实践着古人“诲人不倦”的遗训。这个特点使本书能够成为不少读者案头床边的常备读物，有空看看，常能有新的思考，有更深的理解和收获。

信笔写来，已经有好几页了。即使读者有兴趣看序言，也该去看书中更有趣的内容并开始思考了吧。就此打住。祝愿作者精益求精，根据读者反映和自己的思考发展不断丰富改进本书；更希望早日有新作问世。

张景中

2012 年 4 月 29 日

序二

欣闻《思考的乐趣：Matrix67 数学笔记》即将出版，应作者北大中文系的数学侠客顾森的要求写个序。我非常荣幸也非常高兴做这个命题作业。记得几个月前，与顾森校友及图灵新知丛书的编辑朋友们相聚北大资源楼喝茶谈此书的出版，还谈到书名等细节。没想到图灵的朋友出手如此之快，策划如此到位。在此也表示敬意。我本人也是图灵新知丛书的粉丝，看过他们好几本书，比如《数学万花筒》、《数学那些事儿》、《历史上最伟大的 10 个方程》等，都很不错。

我和顾森虽然只有一面之缘，但好几年前就知道并关注他的博客了。他的博客内容丰富、有趣，有很多独到之处。诚如一篇关于他的报道所说，在百度和谷歌的搜索框里输入 matrix，搜索提示栏里排在第一位的并不是那部英文名为 Matrix（《黑客帝国》）的著名电影，而是一个名为 matrix67 的个人博客。自 2005 年 6 月开博以来，这个博客始终保持更新，如今已有上千篇博文。在果壳科技的网站里（这也是一个我喜欢看的网站），他的自我介绍也很有意思："数学宅，能背到圆周率小数点后 50 位，会证明圆周率是无理数，理解欧拉公式的意义，知道四维立方体是由 8 个三维立方体组成的，能够把直线上的点和平面上的点一一对应起来。认为生活中的数学无处不在，无时不影响着我们的生活。"

据说，顾森进入北大中文系纯属误打误撞。2006 年，还在念高二的他代表重庆八中参加了第 23 届中国青少年信息学竞赛并拿到银牌，获得了保送北大的机会。选专业时，招生老师傻了眼：他竟然是个文科生。为了专业对口，顾森被送入了中文系，学习应用语言学。

虽然身在文科，他却始终迷恋着数学。在他看来，数学似乎无所不能。对于用数学来解释生活，他持有一种近乎偏执的狂热——在他的博客上，油画、可乐罐、选举

制度、打出租车，甚至和女朋友在公园约会，都能与数学建立起看似不可思议却又合情合理的联系。这些题目，在他这本新书里也有充分体现。

近代有很多数学普及家，他们不只对数学有着较深刻的理解，更重要的是对数学有着一种与生俱来的挚爱。他们的努力搭起了数学圈外人和数学圈内事的桥梁。

这里最值得称颂的是马丁·伽德纳，他是公认的趣味数学大师。他为《科学美国人》杂志写趣味数学专栏，一写就是二十多年，同时还写了几十本这方面的书。这些书和专栏影响了好几代人。在美国受过高等教育的人（尤其是搞自然科学的），绝大多数都知道他的大名。许多大数学家、科学家都说过他们是读着伽德纳的专栏走向自己现有专业的。他的许多书被译成各种文字，影响力遍及全世界。有人甚至说他是 20 世纪后半叶在全世界范围内数学界最有影响力的人。对我们这一代中国人来说，他那本被译成《啊哈，灵机一动》的书很有影响力，相信不少人都读过。让人吃惊的是，在数学界如此有影响力的伽德纳竟然不是数学家，他甚至没有修过任何一门大学数学课。他只有本科学历，而且是哲学专业。他从小喜欢趣味数学，喜欢魔术。读大学时本来是想到加州理工去学物理，但听说要先上两年预科，于是决定先到芝加哥大学读两年再说。没想到一去就迷上了哲学，一口气读了四年，拿了个哲学学士。这段读书经历似乎和顾森有些相似之处。

当然，也有很多职业数学家，他们在学术生涯里也不断为数学的传播做着巨大努力。比如英国华威大学的 Ian Stewart。Stewart 是著名数学教育家，一直致力于推动数学知识走通俗易懂的道路。他的书深受广大读者喜爱，包括《数学万花筒》、《数学万花筒 2》、《上帝掷骰子吗?》、《更平坦之地》、《给青年数学家的信》、《如何切蛋糕》等。

回到顾森的这本书上。书的很多章节题目都很吸引人，比如“数学之美”、“几何的大厦”、“精妙的证明”。书的特点就是将抽象、枯燥的数学知识，通过创造情景深入浅出地展现出来，让读者在愉悦中学习数学。比如“概率论教你说谎”、“找东西背后的概率问题”、“统计数据的陷阱”等内容，就是利用一些趣味性的话题，一方面可以轻松地消除读者对数学的畏惧感，另一方面又可以把概率和统计的原始思想糅合在这些小段子里。



数学是美丽的。对此有切身体会的陈省身先生在南开的时候曾亲自设计了“数学之美”的挂历，其中 12 幅画页分别为复数、正多面体、刘徽与祖冲之、圆周率的计算、数学家高斯、圆锥曲线、双螺旋线、国际数学家大会、计算机的发展、分形、麦克斯韦方程和中国剩余定理。这是陈先生心目中的数学之美。我的好朋友刘建亚教授有句名言：“欣赏美女需要一定的视力基础，欣赏数学美需要一定的数学基础。”此书的第二部分“数学之美”就是要通过游戏、图形、数列等浅显概念让有简单数学基础的读者朋友们也能领略到数学之美。

我发现顾森的博客里谈了很多作图问题，这和网上大部分数学博客不同。作图是数学里一个很有意思的部分，历史上有很多相关的难题和故事（最著名的可能是高斯 19 岁时仅用尺规就构造出了正 17 边形的故事）。本书的第三部分专门讲了“尺规作图问题”、“单规作图的力量”、“火柴棒搭成的几何世界”、“折纸的学问”、“探索图形剪拼”等，愿意动动手的数学爱好者绝对会感到兴奋。对于作图的乐趣和意义，我想在此引用本人在新浪微博上的一个小段子加以阐述。

学生：“咱家有的是钱，画图仪都买得起，为啥作图只能用直尺和圆规，有时还只让用其中的一个？”

老师：“上世纪有个中国将军观看学生篮球赛。比赛很激烈，将军却慷慨地说，娃们这么多人抢一个球？发给他们每人一个球开心地玩。”

数学文化微博评论：生活中更有意思的是战胜困难和挑战所赢得的快乐和满足。

书的最后一部分命名为“思维的尺度”，“俄罗斯方块可以永无止境地玩下去吗？”、“比无穷更大的无穷”、“无以言表的大数”、“不同维度的对话”等话题一看起来就很有意思，作者试图通过这些有趣的话题使读者享受数学概念间的联系、享受数学的思维方式。陈省身先生临终前不久曾为数学爱好者题词：“数学好玩。”事实上顾森的每篇文章都在向读者展示数学确实好玩。数学好玩这个命题不仅对懂得数学奥妙的数学大师成立，对于广大数学爱好者同样成立。

见过他本人或看过他的相片的人一定会同意顾森是个美男子，有阳刚之气。很高兴看到这个英俊才子对数学如此热爱。我期待顾森的书在不久的将来会成为畅销书，也期待他有一天会成为马丁·伽德纳这样的趣味数学大师。

汤涛

《数学文化》期刊联合主编

香港浸会大学数学讲座教授

2012.3.5

前　　言

依然记得在我很小的时候，母亲的一个同事考了我一道题：一个正方形，去掉一个角，还有多少个角？记得当时我想都没想就说："当然是三个角。"然后，我知道了答案其实应该是五个角，于是人生中第一次体会到顿悟的快感。后来我发现，其实在某些极端情况下，答案也有可能是四个角或者三个角。我由衷地体会到了思考的乐趣。

从那时起，我就疯狂地爱上了数学，为一个个漂亮的数学定理和巧妙的数学趣题而倾倒。我喜欢把我搜集到的东西和我的朋友们分享，将那些恍然大悟的瞬间继续传递下去。

2005 年，博客逐渐兴起，我终于找到了一个记录趣味数学点滴的完美工具。2005 年 7 月，我在 MSN 上开办了自己的博客，后来几经辗转，最终发展成了一个独立网站 http://www.matrix67.com。几年下来，博客里已经累积了上千篇文章，订阅人数也增长到了五位数。

在博客写作的过程中，我认识了很多志同道合的朋友。2011 年初，我有幸认识了图灵公司的朋友。在众人的鼓励下，我决定把我这些年积累的数学话题整理成册，与更多的人一同分享。我从博客里精心挑选了一系列初等而有趣的文章，经过大量的添删和修改，有机地组织成了五个相对独立的部分。如果你是刚刚体会到数学之美的中学生，这本书会带你进入一个课本之外的数学花园；如果你是奋战在技术行业前线的工程师，这本书或许能不断给你带来新的灵感；如果你并不那么喜欢数学，这本书或许会逐渐改变你的看法……不管怎样，这本书都会陪你走过一段难忘的数学之旅。

在此，特别感谢张晓芳为本书手绘了很多可爱的插画，这些插画让本书更加生动、活泼。感谢明永玲编辑、杨海玲编辑、朱巍编辑以及图灵公司所有朋友的辛勤工作。

同时，感谢张景中院士和汤涛教授给我的鼓励、支持和帮助，也感谢他们为本书倾情作序。

在写这本书时，我在 Wikipedia（http://www.wikipedia.org）、MathWorld（http://mathworld.wolfram.com）和 CutTheKnot（http://www.cut-the-knot.org）上找到了很多有用的资料。文章中很多复杂的插图都是由 Mathematica 和 GeoGebra 生成的，其余图片则都是由 Paint.NET 进行编辑的。这些网站和软件也都非常棒，在这里也表示感谢。

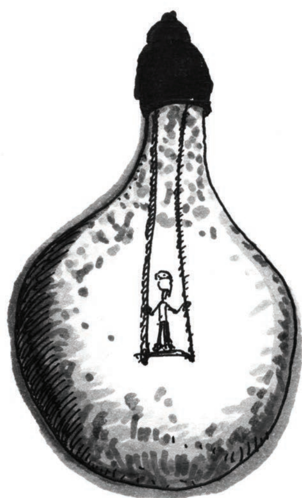
目　　录

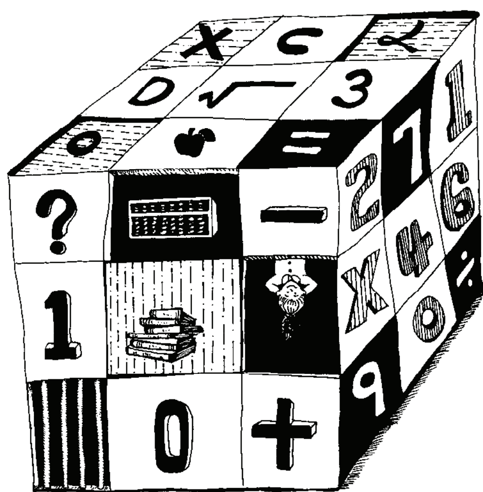




第一部分
生活中的数学

社会学是应用心理学，心理学是应用生物学，生物学是应用化学，化学是应用物理，物理是应用数学。虽然生活的变量太多，建立完整的数学模型几乎是一个不可能完成的任务，但数学玩家们仍然乐此不疲地尝试着用自己的方式理解生活。





1. 概率论教你说谎



在北大念本科时，宿舍里的几个哥们儿特别喜欢玩电脑游戏。M 同学是宿舍里绝对的游戏高手，我们总是被他虐得死去活来的。有段时间，他突然手感不佳，老是发挥失常，反被我们打得狼狈不堪。某天晚上，我们正想继续蹂躏 M 同学，但找遍宿舍楼竟也没发现他的影子。于是我们推测，这家伙肯定到校外的网吧里通宵练技术去了。

第二天一大早，M 同学果然满脸倦意地回到了宿舍。我们几个早有准备，一行人走过去开始拷问他："嘿嘿，昨晚干啥了？" 本以为 M 同学会支支吾吾答不上话来，殊不知他义正词严地答道："我陪女朋友去看通宵电影了。" 我们几个人不服气，问他：



“那电影票呢？”谁知他说了一句“忘了放哪儿了”后，还真煞有介事地在包里翻来翻去。一群人大笑着说：“唉呀，你就别装了吧。“两分钟后，我们全都傻了眼——M同学还真摸出两张电影票。一哥们儿猛地拍了一下M同学的肩膀说：“唉呀，为了骗过我们，你真是煞费苦心啊，居然到影院门口找散场观众买了两张票根!”

笑过之后，我突然开始想，假如M同学为了掩饰自己的丢人行径，真的准备好了伪证的话，那他的演技可不是一般地高明。试着想象以下两个画面。

(1) 几个人不服气，问他：“那电影票呢？”M同学不急不慢地从口袋里掏出两张电影票说：“在这儿呢。”

(2) 几个人不服气，问他：“那电影票呢？”M同学假装到处寻找电影票，过了两分钟才翻出来。

显然，第二种做法更令人相信他真的跑去看通宵电影了。事实上，M同学还能做得更好。

(3) 几个人不服气，问他：“那电影票呢？”M同学条件反射式地说：“电影票早就扔了。”我们继续追问：“不会吧，跟女朋友一起看的电影票就这样扔了，不是你的作风啊。”M同学继续狡辩：“电影票真没了，是不小心搞丢的……”半个小时后，M同学终于（装作）妥协了，说：“那你们看了电影票不要笑我哦。”于是，他（假装）不好意思地交出电影票。我们接过来一看，然后指着他大笑：“你居然和女朋友一起去看《建国大业》？！还是爱国电影通宵连映？!”

这个效果绝对一流，估计我们几乎百分之百地会相信他是真的去看电影了。事实上，很多电影和小说中也有类似的情节，比如《达芬奇密码》中爵士以隐私权为由拒绝警方进入飞机搜查，而事实上警方强行进入后却发现飞机里根本没有别人。爵士事先让大伙儿撤离飞机，并在警方要求搜查飞机时故意造成飞机里还有别人的假象，这样为什么就会让人更加相信爵士反而没有隐瞒什么呢？有趣的是，从概率论的角度来说，这个直觉思维有一个很具有启发性的科学解释。

在概率论中，在知道事件 B 已经发生的情况下，事件 A 发生的概率就记做 $P(A|B)$，它应该等于 $\frac{P(A\cap B)}{P(B)}$，即 A 和 B 同时发生的概率除以 B 本身发生的概率。



例如，投掷一颗骰子，如果已经知道它的点数不超过 3，那么这个点数是奇数的概率就应该等于$\frac{2}{6}$除以$\frac{3}{6}$，即$\frac{2}{3}$。而上述公式中的$P(A\cap B)$又可以等于$P(B|A)\cdot P(A)$，因此我们得到公式$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}$。这个公式叫做贝叶斯（Bayes）定理，它的直观意义就是，当你获知了一个新的信息后，你对原事件的看法所做的改变。若令事件 A 等于“M 同学昨晚在外通宵修炼”，事件 B 等于“M 同学有电影票”，让我们来看看公式中的各个概率的意义。

$P(A)$：M 同学昨晚在外通宵修炼的概率

$P(B)$：M 同学手中有电影票的概率

$P(A|B)$：当 M 同学手中的电影票被发现后，他昨晚在外通宵修炼的概率

$P(B|A)$：如果昨晚 M 同学真的在外通宵修炼，他手中会有电影票的概率

其中$P(A|B)$就是当事人提供了新的证据之后人们对原事件发生概率的看法。利用贝叶斯定理$P(A|B)=\frac{P(B|A)\cdot P(A)}{P(B)}$，我们发现，$P(A|B)$与$P(B|A)$和$P(A)$成正比，与$P(B)$成反比。因此，为了让人们相信事件 A 没有发生，作为伪证的事件 B 一定要具有这样的性质：它本来很可能发生，但伴随着事件 A 一起发生就很不可思议了。通宵电影票就具有这样的性质：有一张通宵电影票根并不罕见，罕见的就是昨晚修炼一夜之后还有一张通宵电影票。为了充分利用这个伪证，让$P(A|B)$变得更低，我们可以从以下三个方面入手。

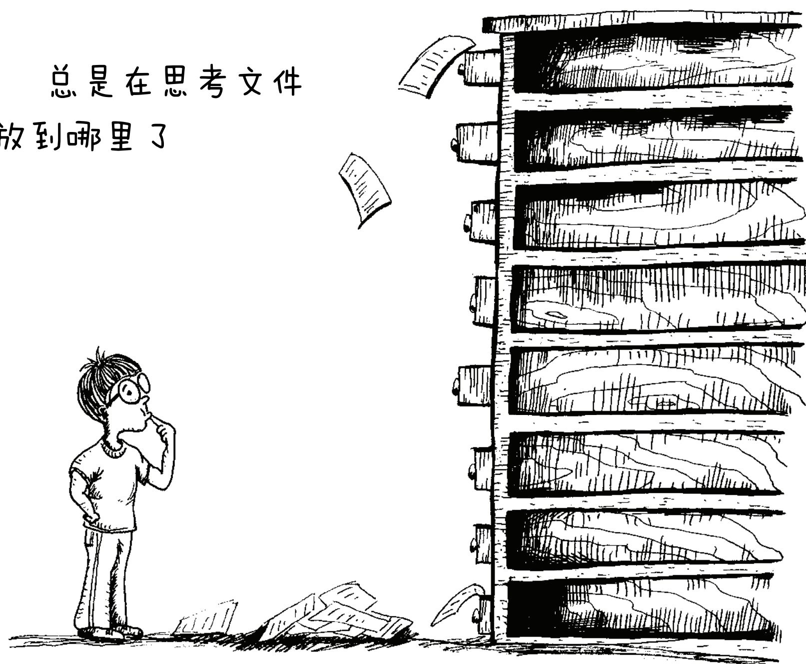
减小$P(B|A)$：不要轻易拿出证据（正如前面所说的策略）。故意做出没法给出证据的样子，让人越来越坚信在事件 A 发生后还能给出证据 B 的概率有多么小。

增加$P(B)$：平时做好铺垫工作。长期保存电影票根，经常提起自己保留纪念物的喜好，让人们相信证据本身的存在并不是什么怪事。

减小$P(A)$：不要一副鬼鬼祟祟的样子，努力提高自己在别人心目中的人品，不至于让人一看见你就说你是不是昨晚又干了坏事。



2. 找东西背后的概率问题



各种违反常理的错觉图片和数学事实告诉我们，直觉并不可靠。其实这本身就是一种错觉，它让我们觉得直觉总是不可信的。而事实上，多数情况下直觉都是可信的，前一节的故事便是一例。我们来看另外一个有趣的例子。

我的书桌有 8 个抽屉，分别用数字 1 到 8 编号。每次拿到一份文件后，我都会把这份文件随机地放在某一个抽屉中。但我非常粗心，有 $\frac{1}{5}$ 的概率会忘了把文件放进抽屉里，最终把这个文件搞丢。

现在，我要找一份非常重要的文件。我将按顺序打开每一个抽屉，直到找到这份文

件为止（或者很悲剧地发现，翻遍了所有抽屉都没能找到这份文件）。考虑下面三个问题。

(1) 假如我打开了第一个抽屉，发现里面没有我要的文件。这份文件在其余 7 个抽屉里的概率是多少？

(2) 假如我翻遍了前 4 个抽屉，里面都没有我要的文件。这份文件在剩下的 4 个抽屉里的概率是多少？

(3) 假如我翻遍了前 7 个抽屉，里面都没有我要的文件。这份文件在最后一个抽屉里的概率是多少？

继续往下看之前，大家不妨先猜一猜，这三个概率值是越来越大还是越来越小？

事实上，三个概率值分别是$\frac{7}{9}$、$\frac{2}{3}$和$\frac{1}{3}$。可能这有点出人意料，这个概率在不断地减小。但设身处地地想一下，这也不是没有道理的。这正反映了我们实际生活中的心理状态，与我们的直觉完全相符：假如我肯定我的文件没搞丢，每次发现抽屉里没有我要的东西时，我都会更加坚信它在剩下的抽屉里；但如果我的文件有可能搞丢了，那每翻过一个抽屉但没找到文件时，我都会更加慌张。我会越来越担心，感到希望越来越渺茫，直到自己面对着第 8 个抽屉，忐忑地怀着最后一丝希望，同时心里想：完了，这下可能是真丢了。

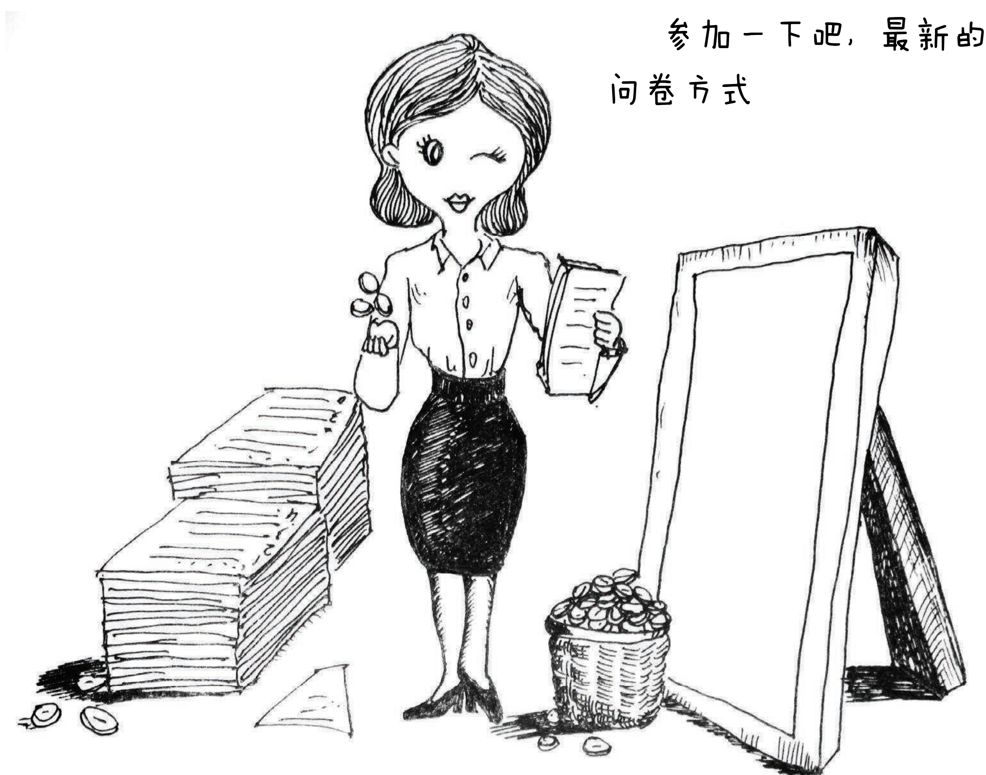
有一个非常巧妙的方法可以算出上面三个概率值来。

注意到，平均每 10 份文件就有两份被搞丢，其余 8 份平均地分给了 8 个抽屉。假如我把所有搞丢了的文件都找了回来，那么它们应该还占 2 个抽屉。这让我们想到了这样一个有趣的思路：在这 8 个抽屉后加上 2 个虚拟抽屉——抽屉 9 和抽屉 10，这两个抽屉专门用来装我丢掉的文件。我们甚至可以把题目等价地变为：随机把文件放在 10 个抽屉里，但找文件时不允许打开最后 2 个抽屉。当我已经找过 n 个抽屉但仍没找到我想要的文件时，文件只能在剩下的 $10-n$ 个抽屉里，但我只能打开剩下的 $8-n$ 个抽屉，因此所求的概率是 $\frac{8-n}{10-n}$。当 n 分别等于 1、4、7 时，这个概率值分别是$\frac{7}{9}$、$\frac{2}{3}$和$\frac{1}{3}$。

如果把$\frac{8-n}{10-n}$写成$1-\frac{2}{10-n}$，就很容易看出，当$0 \leqslant n \leqslant 8$时，它是一个递减函数。



3. 设计调查问卷的艺术



设计一张合理的调查问卷并不是一件容易的事情，需要综合考虑各方面的因素。比方说，假如你需要在调查表中问一个极度隐私的问题，尽管在调查表上再三强调你们的保密措施，但你真的指望所有人都能够如实地回答吗？你真的指望会有人在“我有外遇”前面打一个勾，然后把表递到问卷回收人的手中吗？

有什么方案能够从理论上保证个人隐私绝对不可能被泄露，让每个人都能够放心地填写，并且问卷回收之后能够得到一个准确的统计结果呢？为了方便起见，假设这个问题的答案只有“是”和“否”两个选项。

这里提供一个很漂亮的解决方案。在问卷上要求每个人准备一枚硬币（或者叫问卷发放人给每个人发一枚一元钱的硬币，顺便也当做填写问卷的酬谢）。对于指定的隐私题目，请填写人投掷一次硬币：如果正面朝上，则如实填写个人的真实情况；如果反面朝上，那么就再投掷一次硬币，正面就选“是”，反面就选“否”。当然，若第一次投掷硬币为正的话，填写人完全可以假装再投一次硬币来掩人耳目。这样，别人永远不知道你在“我有外遇”前面打了勾是因为你真的有婚外恋，还是因为那个答案是投掷出来的。

回收所有的问卷后，我们需要推测出，在那些如实回答了问题的人中，有多少人选择了“是”。假设回收到的有效问卷有 m 份，其中该问题答“是”的有 n 个人。那么，如实回答了该问题的人平均有 $\frac{m}{2}$ 个。另外 $\frac{m}{2}$ 人则是抛币作答的，其中有大约 $\frac{m}{4}$ 的人“被迫”答了“是”。因此，我们所需要的最终结果就是 $\frac{n-m/4}{m/2}$。

把这个算法写在问卷上，让大家知道问卷调查结果将如何统计，以便让大家严格遵守该问题的填写方法。



4. 统计数据的陷阱

和统计数据打的交道多了，什么见鬼的事情都能遇上。统计数据显示，在铀矿工作的工人居然与其他人的寿命相当，有时甚至更长！难道统计结果表明在铀矿工作对身体无害么？

当然不是！其实，统计数据本身并没有说谎，铀矿工人的寿命真的不比普通人低，难就难在我们如何拨开数据的外表，从中挖掘出正确的信息。事实上，只有那些身强体壮的人才会去铀矿工作，他们的寿命本来就长一些，正是因为去了铀矿工作，才把他们的寿命拉低到了平均水平，造成了数据的“伪独立性”。这种现象常常被称为“健康工人效应”。

类似地，有数据表明打太极拳的人和不打太极拳的人平均寿命相同。事实上呢，太极拳确实可以强身健体、延长寿命，但打太极拳的人往往是体弱多病的人，这一事实也给统计数据带来了虚假的独立性。

有虚假的独立性数据，就有虚假的相关性数据。统计数据显示，去救火的消防员越多，火灾损失越大。初次听到这样的结论，想必大家的反应都一样：这怎么可能呢？仔细想想你就明白了：正因为火灾损失大，才会有很多人去救火。因果关系弄颠倒了。数据只能显示两件事情有相关性，但并不能告诉你它们内部的逻辑关系。

事实上，两个在统计数据上呈现相关性的事件，有可能根本就没有因果关系。统计数据表明，冰淇淋销量增加，鲨鱼食人事件也会同时增加。但这并不意味着，把冰淇淋销售点全部取缔了，就能减小人被鲨鱼吃掉的概率。真实的情况则是，这两个变量同时增加只不过是因为夏天来了。统计数据显示，足球队的获胜率，竟然与队员的球袜长度成正比。难道把队员的球袜都换长一些，就能增加进球数了吗？显然不是。数据背后真正的因果关系是，球队的获胜率和队员的球袜长度都与队员的身高呈正相



关，这导致了获胜率与球袜长度之间表现出虚假的相关性。

类似的例子还有很多。统计数据表明，手指越黄的人，得肺癌的概率越大。但事实上，手指的颜色和得肺癌的概率之间显然没有直接的因果联系。那么为什么统计数据会显示出相关性呢？这是因为手指黄和肺癌都是由吸烟造成的，于是又营造出一种虚假的相关性。

读到这里，大家脑子里或许会产生这么一个颠覆性的念头：根据同样的道理，我们又凭什么说吸烟会致癌呢？万一吸烟和肺癌也都是由另外一个东西同时导致的怎么办？

其实，要想知道吸烟与癌症之间究竟是否有因果联系，方法本来很简单：找一群人随机分成两组，规定一组抽烟一组不抽烟，十几年后再把这一拨人找回来，数一数看是不是抽烟的那一组人患肺癌的更多一些。这个实验方法本身是无可挑剔的，但它太不道德了，因此我们只能考虑用自然观察法，选择一些本来都不吸烟的健康人进行跟踪观察，然后呢，过一段时间这拨人里总会出现一些失意了、堕落了犯上烟瘾的人，于是随着时间的流逝这帮人自然而然地分成了可供统计观察的两组人。注意，这里"是否吸烟"这一变量并不是通过随机化得来的，它并没有经过人为的干预，而是自然区分出来的。这是一个致命的缺陷！统计结果表明，犯上烟瘾的那些人得肺癌的几率远远高于其他人。这真的能够说明吸烟致癌吗？仔细想想你会发现这当然不能！原因恰似之前提过的例子：完全有可能是因果关系颠倒了，或者某个第三方变量同时对"爱吸烟"和"患肺癌"产生影响。1957 年，费希尔（Fisher）提出了两个备选理论：癌症引起吸烟（烟瘾是癌症早期的一个症状），或者存在某种基因能够同时引起癌症和烟瘾。

现实中的统计数据往往会表现出一些更加诡异复杂的反常现象，带来更多意想不到的麻烦。辛普森（Simpson）悖论是统计学中最有名的悖论：各个局部表现都很好，合起来一看反而更差。统计学在药物实验中的应用相当广泛，每次推出一种新药，我们都需要非常谨慎地进行临床测试。但有时候，药物实验的结果会匪夷所思。假设现在我们有一种可以代替安慰剂的新药。统计数据表明，这种新药的效果并不比安慰剂好：



	有　　效	无　　效	总 人 数
新药	80	120	200
安慰剂	100	100	200

简单算算就能看出，新药只对 40%的人有效，而安慰剂则对 50%的人有效。新药按理说应该更好啊，那问题出在哪里呢？是否因为这种新药对某一类人有副作用？于是研究人员把性别因素考虑进来，将男女分开来统计：

	男性有效	男性无效	女性有效	女性无效
新药	35	15	45	105
安慰剂	90	60	10	40

大家不妨实际计算一下：对于男性来说，新药对高达 70%的人都有效，而安慰剂则只对 60%的人有效；对于女性来说，新药对 30%的人都有效，而安慰剂则只对 20%的人有效。滑稽的一幕出现了：我们惊奇地发现，新药对男性更加有效，对女性也更加有效，但对整个人类则无效！

这种怪异的事屡见不鲜。曾有一个高中的师弟给我发短信，给了我两所大学的名字，问该填报哪个好。我考虑了各方面的因素，甚至非常认真地帮他查了一下两所大学的男女生比例，并且很细致地将表格精确到了各个院系。然后呢，怪事出现了：A 学校的每个院系的女生比例都比 B 学校的同院系要高，但合起来一看就比 B 学校的低。当然，进错了大学找不到女朋友是小事，但医药研究需要的是极其精细的统计实验，稍微出点差错的话害死的可就不是一两个人了。

上面的例子再次告诉我们，统计实验的“随机干预”有多么重要。从上面的数据里我们直接看到，这个实验的操作本身就有问题：新药几乎全是女性在用，男性则大都在用安慰剂。被试者的分组根本没有实现完全的随机化，这才导致了如此混乱的统计结果。不难设想，如果每种药物的使用者都是男女各占一半，上述的悖论也就不会产生了。当然，研究人员也并不笨，这么重大的失误一般还是不会发生的。问题很可能出在一些没人注意到的小细节上。比如说，实验的时候用粉色的瓶子装新药，用蓝色的瓶子装安慰剂，然后让被试人从中随机选一个来用。结果呢，女孩子们喜欢粉色，选的都是新药；男的呢则大多选择了蓝瓶子，用的都是安慰剂。最后，200 份新药和 200 份安慰剂正好都发完，因此不到结果出来时，就没有人会注意到这个微小的性



别差异所带来的统计失误。

当然，上面这个药物实验的例子并不是真实的，一看就知道那个数据是凑出来方便大家计算的。不过，永远不要以为这种戏剧性的事件不会发生。《致命的药物》一书详细披露了 20 世纪美国的一次重大药害事件，其原因可以归结到药物实验上去。人们推测，事故发生的原因就与一些类似的统计学现象相关。

这些离奇的统计学现象有时会让人感到恐慌：连统计数字也不可靠了，还有什么能真实地反映这个世界运转的规律呢？



5. 为什么人们往往不愿意承担风险?

张志强的博客“阅微堂”[①]也是一个非常有名的数学博客。他曾经在博客中写过，有两门课程是所有大学生都应该学习的，一是概率论，二是经济学，这两门课分别代表着两种生活中的思维方式。我非常赞同这个观点。概率论并不仅仅是一门关于概率的学问，它把世间发生的一切抽象为“事件”，把那些充满不确定性的复杂机制抽象为一个个“随机过程”，给我们带来了一种全新的世界观。同样地，经济学也并不是一门关于经济的学问。利用经济学模型，我们可以解释人们日常生活中很多看似不合情理的决策。

假设现在有两份临时工作供你选择，它们的工作内容都完全相同，只是报酬方式不一样：工作一，有 $\frac{1}{2}$ 的概率获得 500 元，有 $\frac{1}{2}$ 的概率获得 1500 元；工作二，百分之百地稳拿 1000 元钱。虽然看上去两种选择的平均收入都一样，但是人们往往更愿意选择后一份工作，尽可能避免前一种工作的风险。为什么面对期望收入相同的事件，人们往往愿意选择风险更小的那一个呢?

关键的原因在于，收入本身并不重要，我们关心的是它能带给我们的好处，或者说它给我们带来的幸福感、满足感。在经济学中，我们用“效用”这个词来表示这种主观上对收益的评估结果。

这里，我们有一个重要的假设：收入的边际效用（收入每增加一个单位所带来的额外效用）是递减的。换句话说，增加同样多的收入，低收入者主观上会感觉自己收益了很多，本来就是高收入的人则觉得这点儿收入算不了什么。人们往往会觉得，收入从 100 元增加到 200 元所带来的效用，要远远大于收入从 800 元增加到 900 元所带来的效用。因此，如果把个人收入和它给人带来的效用画成一条曲线的话，大致就如

① 博客地址为 http://zhiqiang.org/blog。

图 1 中的那条曲线。

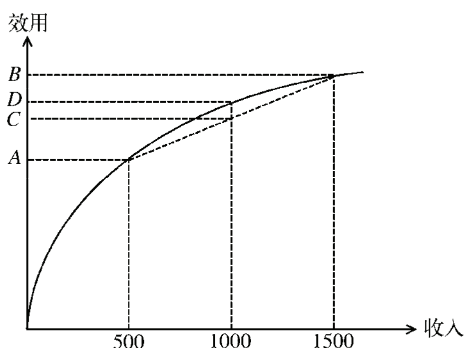


图　1

假如你获得了 500 元钱，你所得到的效用就用 A 点来表示；假如你获得了 1500 元，你所得到的效用就在 B 点。因此，工作一带给你的平均效用就用 A 和 B 的中点 C 来表示。但是，如果直接就给你 1000 元，你将会得到一个大于 C 的效用 D。这表明，直接选择工作二所带来的效用要高于工作一带给你的平均效用，自然人们都会选择工作二了。

因此，经济学中有这样一个定理：如果一个人认为自己收入的边际效用是递减的，那么这个人就是一个风险规避者。对于期望收入相同的两件事来说，他愿意去做风险更小的那一件。

事实上，风险规避者甚至有可能通过减少自己的收入来避免可能的风险。在图 2 中我们可以看到，如果工作二所提供的稳定收入值高于 x 元，风险规避者就会毫不犹豫地选择工作二，即使它的收入低于工作一的平均收入。也就是说，一个风险规避者愿意花费 $1000-x$ 元钱来避免他可能面对的风险。

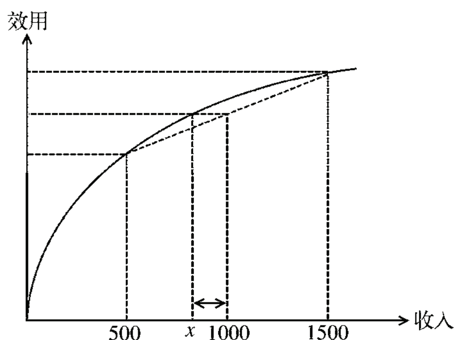


图　2



6. 消费者承担消费税真的吃亏了吗？





其实，我本来对经济学不感兴趣。一次偶然的机会，我在朋友的寝室里看到了传说中经济学最经典的教材之一——曼昆（Mankiw）的《经济学原理》。好奇心驱使我随手翻开了一页，读了一小段与征税有关的讲义，于是立即爱上了经济学，并且果断选修了微观经济学的课程。这是我大学期间收获最大的课程之一。

可能是因为个人的经历吧，我觉得征税问题特别适合用作微观经济学的入门话题。因此，我准备在这里复述一下《经济学原理》中的这段讲义，希望从未接触过经济学的数学爱好者们能够喜欢上这门学问。

我打算偷一个小懒，直接用原书上的例子——冰淇淋。

与众多其他市场一样，冰淇淋市场的需求曲线与供给曲线的走向是正好相反的。当冰淇淋的价格增加时，越来越多的消费者觉得吃冰淇淋的享受不值这么多钱，从而退出了消费市场，于是市场的总需求量越来越低。反之，冰淇淋的价格越低，能够提供冰淇淋的生产商也越少，因为越来越多的卖者认为他们没有赚头，从而退出市场竞争。两条曲线有一个交点，这个交点叫做市场均衡。对应的价格叫做市场均衡价格，对应的数量则叫做均衡数量（见图 1）。在均衡价格下，买者的需求与卖者的供给数量正好相当，市场上的每个人都得到了满足。若市场价不等于均衡价格时，供给数量和需求数量将不再平衡；供不应求将导致价格上涨，供大于求则导致价格下跌，最终还是会自发地调整到均衡价格。

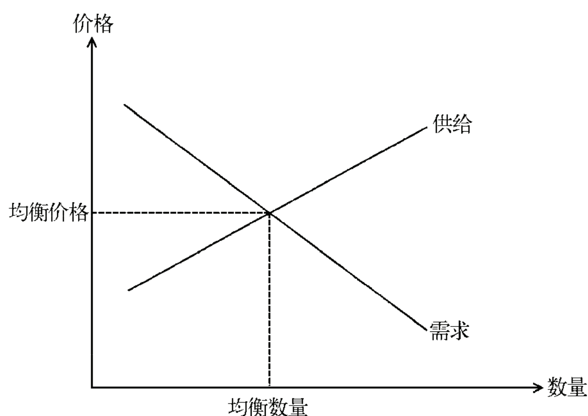


图　1



现在呢，有趣的事情发生了。假设有一个地方具有相当浓厚的冰淇淋文化，该地政府打算举办一个年度冰淇淋节。为了筹到这项活动的经费，政府决定：卖方每卖出一个冰淇淋，政府就向卖者征收 0.5 美元的税。于是，各大冰淇淋制造商上街游行，宣称这个税应该由买者支付。而消费者协会则声援政府，坚持认为这部分税应该由冰淇淋生产商支付。两大游说集团吵成一团。为此，我们不妨仔细研究一下，如果这部分税由消费者来承担的话，会发生什么奇特的事情。

假设政府向消费者征税。消费者自然会觉得自己亏大了：每买一个冰淇淋还要多付 0.5 美元。消费者并不关心市场价格，只关心自己的实际支出，因此，如果原本我能接受 2 美元的冰淇淋，现在我只愿意接受 1.5 美元的了，因为我还得额外支付 0.5 美元的税。换句话说，需求曲线向下移动了 0.5 个单位（见图 2）。新的需求曲线与供给曲线产生了新的交点，市场的均衡数量变少了，市场均衡价格也降低了。假如说，没有征税时市场均衡价格为 3.0 美元，现在的市场均衡价格为 2.8 美元。但消费者要交 0.5 美元的税，因此消费者支付的实际价格是 3.3 美元。我们可以看到，政府若向消费者征税，则卖方损失了 0.2 美元的收益，买方则多付出了 0.3 美元。这 0.5 美元的税实际上是由双方共同承担的。究竟哪一边分担得多些是由两条线的斜率决定的。

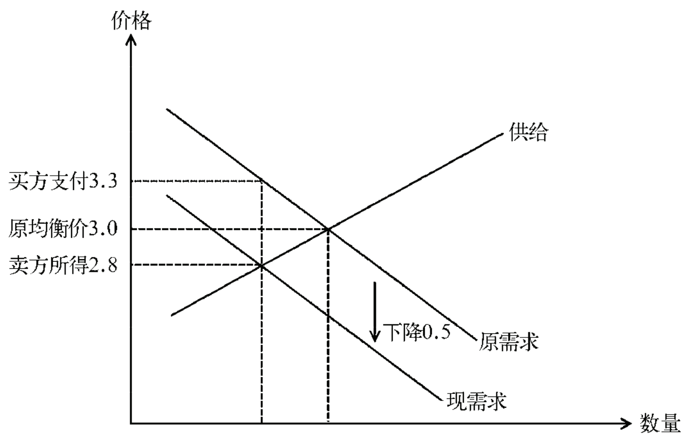


图　2

20 世纪美国曾经大规模地向消费者征收奢侈品消费税。因为政府觉得，买奢侈品的都是富人，因此对奢侈品征收消费税其实是非常巧妙地变相向富人多征一些税。殊

不知，奢侈品不是生活必需品，只要价格抬高一点，便有大量的消费者退出市场，反正有的是地方花钱，买点房子啊，出去旅游啊，要实在得多。反过来，奢侈品的供给曲线则非常地陡，即使价格变化很大，产量变化仍然不大，毕竟生产制造奢侈品需要用很多时间、人力和设施，这些既定因素使得生产商无法快速应对市场需求变化。可见，需求曲线比供给曲线要“平”得多。结果呢，明明是向买方征税，税反而几乎都由生产者承担；而这些生产者并不是富人，奢侈品税的重担落在了中产阶级身上。政府的决策适得其反。

别着急，冰淇淋的故事还没讲完呢。我们再来看看，如果果真向生产商征税，结果又如何呢？显然，生产者必然会觉得自己亏了，原本可以卖 2 美元，现在卖了后只能得 1.5 美元了。因此，为了弥补这 0.5 美元的损失，卖方只接受比原来高 0.5 美元的市场价格。其结果是，供给曲线上升了 0.5 个单位（见图 3），从而使得市场均衡价格从 3.0 美元增加到了 3.3 美元。但这 3.3 美元并不全部归卖方，卖方要交给政府 0.5 美元的税，因此事实上卖方只能得到 2.8 美元。结果呢，向生产者征税的效果与向消费者征税的效果完全一样。

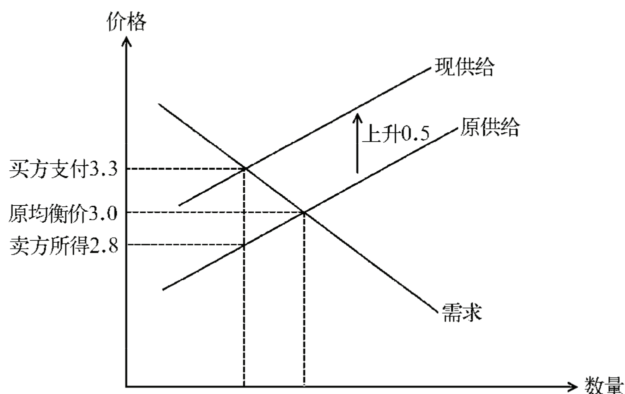


图 3

搞了半天，最开始两边在那里拼了命地争论，结果却完全没有必要——不管向谁征税，结果都是一样的。



7. 价格里的阴谋

很多常见的商品，比如大米和白菜等，它们的买家和卖家都很多，产品本身的差异也不大。因此，个人行为是无法改变整个市场的，价格完全由整个市场的供求决定。这种市场叫做完全竞争市场。在完全竞争市场中，卖家自己是无法操纵价格的。

还有一些产品就不同了。比如铁路和电力等市场，产品的提供商通常只有一个企业，这个企业就能随意调整产品的价格。电信和航空等产业也不是随便哪个人就可以白手起家说干就干的，新企业的参与和旧企业的退出都需要耗费巨大的成本，这也决定了商品的提供商必然不会很多，企业有自主定价的空间；还有衣服、手机和书报等商品，不同商品之间的差异很大，每一种产品都有它的独特性，因此这些行业也不是完全竞争，生产商也有自己定价的权利。由此引发了一个有趣的话题——如何制定价格才能让生产商的利益达到最大呢?

这里有一个两难的问题：价格定得太低，赚不到钱；价格定得太高，没人买。这是传统定价策略的一个巨大的缺陷：不管你把价格定到多少，你都觉得不好——价格再高点或许就能从某些买家手里赚到更多，价格再低点或许就能赢来一些新的买家。要是有办法给愿意高价购买的人卖贵点，给只想便宜买的人卖便宜点的话就好了。这种放弃统一定价，为不同消费者制定不同价格的策略就叫做“价格歧视”。

对于商家来说，最完美的情况就是拥有看透每个买家的读心术，能知晓每个人愿意支付的最高价格，并且抵着这个价格卖给他。这种为每个人“量身定价”的理想情况被称为“一级价格歧视”。在现实生活中，一级价格歧视显然是不大可能发生的。不过有一些例子却非常接近一级价格歧视。比方说小商铺中的讨价还价，最后的成交价格因人而异，这就有点一级价格歧视的味道。聪明的卖家在报价前会先问“你觉得它值多少钱”，目的就在于摸清你的心理价位。对于一些不大会砍价的人，回答卖家



的这个问题几乎就是彻底暴露自己愿意支付的最高价格，于是市场上又诞生了一个悲剧的消费者。

和每个消费者讨价还价虽然很接近梦想中的一级价格歧视，但这并不能在每个行业里都办到。除了“明码标价”等政策上的原因之外，有时候还有一些更直接的原因。比方说电信业，话费和流量费就只能统一定价，与每个消费者都搞讨价还价根本不可能实施，况且消费者众多，费用信息是完全透明的。因此，商家还得绞尽脑汁想点儿别的办法来区分不同档次的消费者才行。

我们就用数据流量费来举例子吧。在 GPRS 服务出现之初，人们用 GPRS 可以干的事情并不多，因此我们假设消费者的需求都差不多。每个月 30MB 的流量对于数据流量的消费者来说已经足够了，再多了也用不上。但是，这 30MB 的流量在消费者心中的价值并不一样。对于一个饿汉来说，第一个烧饼的价值显然比第七个烧饼的价值更高。对于消费者来说，每多 1MB 流量所带来的价值也是递减的。我们假设，为了得到头一兆的流量消费者愿意出 3 元钱，但消费者只愿意再花 2.9 元获得额外的一兆，第三兆则只值 2.8 元钱，等等。我们把消费者对每单位流量的估价用图 1 所示的柱状图表示，所有竖条面积的总和就是这 30MB 的流量在消费者心目中的总价。

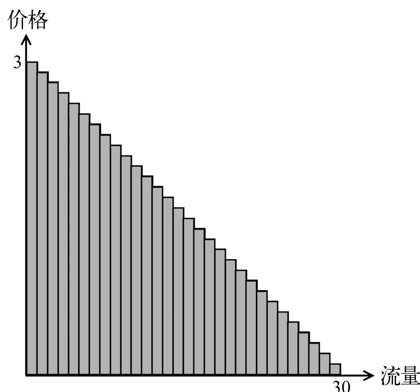


图 1

如图 2 所示，我们近似地用一条斜线来反应流量和价格之间的关系，斜线下方的三角形面积就可以看作是一个消费者为了得到 30MB 愿意支付的总价——约 45 元。



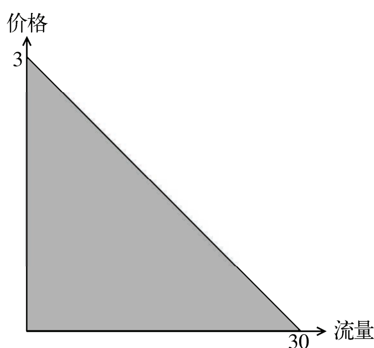


图 2

对于电信公司来说，怎样定价才能赚到更多的钱呢？和上面提到的定价困境一样，流量单价无论怎么设定都不完美。比方说，我们规定每兆的单价为 1 元，于是消费者就会打起如意算盘，算出一个让自己赚得最多的购买数量。结果就是，消费者只愿意购买 20MB，如图 3 所示，因为此时自己的获利减去实际的支出达到最大，每再多买一点就会又亏一点。这样的话，假设提供数据服务的成本为 0，服务提供商也只能赚到一个小矩形区域这么多钱（20 元），斜线下方的其他区域都被放掉了。让这个矩形面积达到最大的方法是把单价定到 1.5 元，这样可以从每个消费者手中赚 22.5 元钱，但获得的利润仍然只有斜线下方面积的 1/2。有没有办法榨干消费者的每一分钱呢？有！那就是放弃按单价收费的办法，直接推出一个 45 元 30MB 的套餐。由于每个消费者购买 30MB 的流量所愿意支付的最高价格恰好也就是 45 元，因此消费者将接受这个价格，于是服务提供商将赚到斜线下方的所有面积。取消按单价收费的办法后，消费者将别无选择，只要套餐价格没超过带给他的价值，他都会去买。为什么电信业务里总是有那么多套餐，秘密也就在这里了。

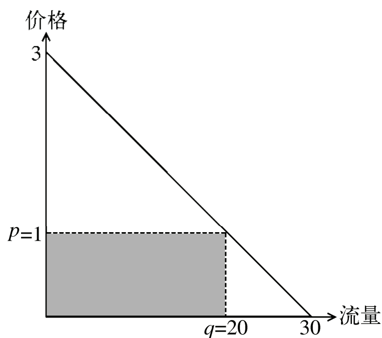


图 3



现在，有趣的问题来了。假设数据流量市场上突然出现了一类新的消费者。或许是由于这类消费者用 GPRS 比较频繁，或许是由于他们用 GPRS 的方式比较费流量，总之 40MB 才能满足他们的需求。他们对每单位流量的价值估算也是随着流量增多而递减的。他们愿意为头一兆流量花费 4 元钱，但只愿意花 3.9 元钱购买第二兆，依此类推。这样的话，市场上就出现了两种消费愿望不同的消费者，我们不妨把他们分别叫做“低端消费者”和“高端消费者”。

若只推出一个 30MB 套餐，则只能赚到两个面积 A 的钱，荒废了高端消费者的巨大潜力；若只推出 40MB 套餐，则只能赚到面积 $A+B+C$ 的钱，完全无视了低端消费者的购买力。为了兼顾两类消费者，从消费者身上榨取出最多的钱，就需要放弃统一定价策略，并同时推出两种套餐：45 元钱 30MB，以及 80 元钱 40MB。如图 4 所示，低端消费者愿意用面积 A 所代表的钱数去购买 30MB，高端消费者愿意用面积 $A+B+C$ 所代表的钱数购买 40MB，因此他们都能接受为自己准备的套餐，以愿意支付的最高价格购买数据服务。这就是鲜活的价格歧视：给不同的消费者制定不同的价格。但此时，我们发现了一个之前不曾遇到过的问题：高端消费者可能会发现，买前一种套餐似乎更划得来——对于高端消费者来说，30MB 的价值等于面积 A 加上面积 B，但现在只需要用面积 A 就能拿到这 30MB，又何乐而不为呢？另外的 10MB 流量对高端消费者的价值只相当于区域 C 的面积，却需要在低端套餐的基础上再加上面积 $B+C$ 的钱才买得到，明显亏了很多。这就是实现价格歧视真正最困难的地方：既然不能靠讨价还价等手段区别消费者，在一个开放的市场环境中，如何阻止高端消费者模仿低端消费者去消费低端套餐呢？

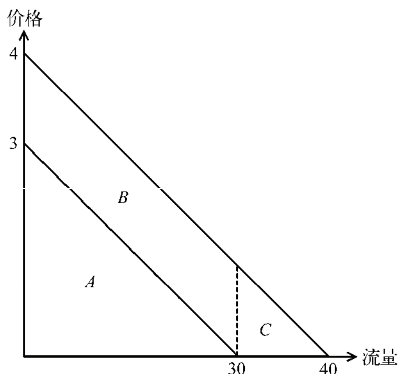


图 4



为了让高端消费者自动去选择高端套餐，我们必须要让高端消费者觉得，购买高端套餐要比购买低端套餐更划得来。因此，我们想到以下这种改进的定价方式。低端套餐是面积 A 购买 30MB，高端套餐是面积 $A+C$ 购买 40MB。高端消费者会发现，在购买了 30MB 的流量之后，再获得额外的 10MB 对他而言的价值相当于面积 C，恰好也就是购买 40MB 套餐的额外付出。因此，高端消费者会觉得多花一个面积 C 的金额是值得的，从而主动去选择后面那一种套餐。这样，服务提供商将从两类消费者中赚取到的总面积为 $A+A+C$。这种套餐定价虽然不能赚到消费者愿意支付的每一分钱，但它能自动把两类消费者区分开来，让每类消费者都会自动选择适合他的套餐，实现了消费者的区别对待，从而赚到比统一定价更多的钱。我们把这种给不同量的商品制定不同的价格，得以让高端消费者自动选择高价位商品的定价策略叫做“二级价格歧视”。这样的例子在生活中很多见。“一件 30 元两件 50 元”和“量大从优”的价格策略本质上都是二级价格歧视的典型例子。

有趣的是，上面这种套餐设置还不是最好的，它还能继续改进。由于高端消费者愿意花费的钱更多一些，我们可以想办法拉大高端套餐和低端套餐的差距，从而向高端消费者收取更高的费用。例如，按照图 5 设置两个套餐，面积 A' 购买 x 兆，面积 $A'+D+E+C$ 购买 40MB。低端消费者会发现他购买前 x 兆愿意支付的钱正好也就是面积 A'，因此愿意接受前一个套餐；高端消费者发现把流量扩充到 40MB 愿意多支付的钱正好也就是面积 $D+E+C$，因此会购买 40MB 套餐。这时，服务提供商赚到的为一个 A' 的面积，加上 $A'+D+E+C$ 的面积，和原来相比少赚了一个 D 的面积，但多赚到了一个 E 的面积。由于区域 E 要比区域 D 大一些，因此这个套餐比原来更好。

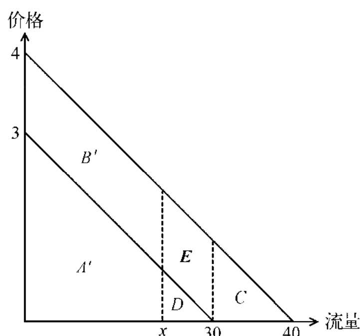


图 5

x 到底取多少才能达到最优呢？注意，只要 D 区域的左边界比 E 区域的左边界更短，把 x 的值减小一点总能保证面积 E 的变化量大于面积 D 的变化量。当 $x=20$ 时，D 、E 两块区域的左边界一样长了，低端套餐的低端化也就到了极限。因此，如图 6 所示，在这个例子中，最终的二级价格歧视策略是设定 20MB 和 40MB 两个套餐，它们的价格分别为面积 A' 和面积 $A'+D+E+C$ 所代表的钱数。

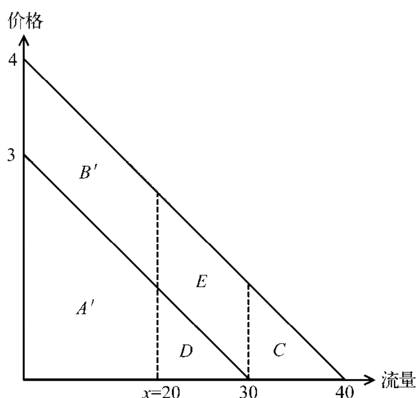


图　6

这里我们看到了一个有趣的现象：让低端产品更低端，反而会增大生产商的收益。只需要注意到例子中的横轴不一定总是代表商品的数量，它也可以用来表示商品的质量，我们就会发现二级价格歧视理论可以解释生活中很多奇怪的现象。联邦快递服务表面上有次日到达、隔日到达、普通到达三种，但显然普通到达的快递并不是真的需要更长的运输时间才能到。在同一天寄出的快递，即使选用了不同的服务，它们显然也都是在同一天到的。只是，隔日到达的快递会在仓库里多囤一天，普通到达的快递则会被搁置更久。有人会想，这不是有毛病吗？为什么明明今天就能送到的东西非要明天才送到？事实上，这种看似很不合理的做法正是前面所说的二级价格歧视。快递公司人为地把快递服务分成了三种不同的档次，有意设置低端服务，从而让消费者根据自己的消费水平对号入座。轮船的四等舱又脏又臭，很多乘客都抱怨，明明只需要很小的成本就能稍微改善一下四等舱的环境，为什么不这么做呢？其实，这也是价格歧视的需要。为了区分出不同档次的产品，商家有意设置了一个低端消费品，供那些支付意愿较低的人购买。



二级价格歧视还有一些更匪夷所思的例子。为了实现价格歧视，产品研发部门有时会面对一些看似不可理喻的设计需求——IBM 研发打印机时曾经专门研发过一种可以减慢打印速度的部件。超市新进一批货物后，往往会举办特卖会贱价销售运输过程中有所损坏的商品。每次运输中真的都会产生那么多不小心受损的东西吗？有人惊奇地发现，这些号称是运输中撞伤的商品竟然都是到货之后商家自己用锤子砸坏的！有了价格歧视理论，生活中的很多怪现象都有了合理的解释。

除了用不同档次的商品来区分消费者，有时候，商家还有其他办法直接区分出消费水平不同的买家。如果商家能够成功区别出不同档次的消费者，无需拐弯抹角，直接就给他们提供不同的价格，这就叫做“三级价格歧视”。游乐园门票、电影票和火车票等商品不大能分出个一等二等，因此二级价格歧视在这儿没有什么用武之地。不过，商家仍然能够想出区别定价的奇招：持有学生证可以享受优惠。由于学生群体消费水平较低，而借助学生证又能轻易将这类消费者区别开来，因此商家可以直接给这一类人提供优惠价，从而既能保证榨取高端消费人群，又不至于损失了低端消费人群。同一件商品在不同省市的价格不同，高速公路对不同车型收取不同的费用，这些都是最典型的三级价格歧视。

当然，还有一些非典型的、很隐蔽的三级价格歧视。商家经常在暗中布置好一盘棋，根据你的行为来分辨你的消费档次。在很多商场、餐厅或者酒店，获取更低折扣的办法竟然就是简单地问一句“打折吗”。别小看这个小细节，问不问这一句话很大程度上就反映出了买家的消费水平。按照这个行为细节把消费者分为两个档次，给他们提供不同的价格，兼顾不同消费人群，这就是相当隐蔽的三级价格歧视策略。电子商务网站也能根据用户操作区别出不同的消费人群。一些阴险狡诈的网站可能会在用户点击“按价格从高到低排序”后有意给出更高的价格，目的就是从高端消费者那里赚到更多的钱。

还有一些更隐蔽的三级价格歧视。优惠券的印刷和发放都需要耗费不少的成本，那麦当劳为什么不直接在餐厅提供折扣，而偏偏要用优惠券的方式提供折扣呢？其实，提供优惠券就是一个非常隐蔽的三级价格歧视。据说，拿到优惠券的人当中，只有 30%的人会有意把它留下来供以后使用，另外 70%的人不是放着放着就弄丢了，就是放着放着就过期了，甚至有很多人拿到优惠券就直接扔掉了。根据这一点，消费者

就自动分为了两个群体。这样，商家便能从高端消费者手中榨取到更多的钱，并为那些对价格很敏感的低端消费者提供优惠价。在国外买很多电子产品时，有一种价格优惠策略叫做“邮寄回扣”，就是说买完东西后把收据、反馈卡、回扣申请表等物品整理好并寄回厂家，厂家就会以支票的形式返赠多少多少钱。返赠的金额少则几十元，多则一百多元，对消费者来说无疑是一个巨大的诱惑。但事实上，申请回扣是一件很麻烦的事情，需要寄回厂家的东西少了任何一样都不行。因此，回了家后真正认真整理回扣申请资料的人并不多，很多人要不就是嫌手续麻烦不弄了，要不就是放着放着就忘了。只有对价格特别敏感，真正在乎回扣的消费者才会花精力去申请回扣。高端消费者和低端消费者就这样区别开了。

为了榨干消费者的每一分钱，除了价格歧视以外，商家还想出了各种招数。一种看上去似乎与此无关的定价策略叫做“两部分定价”。游乐园、酒吧之类的地方广泛存在两部分定价的现象，即在消费者消费之前必须先一次性支付一定数量的“入场费”，入场之后才可以按单价支付你所消费的商品。为什么商家要把费用分成这么两层呢？其实，根本目的还是在于从消费者手中赚到更多的钱。

为了说明为什么两部分定价能赚到更多，我们不妨以游乐园来举例。为了简便起见，我们假设游乐园里只有一个游乐项目，比方说过山车。去游乐园的人只有一个目的，就是去玩过山车。不过，过山车老玩也没意思，随着玩的次数增加，游客获得的“爽感”将逐渐减小，具体地说，第 n 次坐过山车只能给他带来相当于 $100-10n$ 元的价值（这也就是他第 n 次乘坐过山车愿意支付的最高价格）。我们再假设，运营过山车的成本是平均每人次 60 元。那么，游乐园应该怎样定价才能从消费者手中赚到最多的钱呢？

首先注意到，传统定价策略依旧有前面已经讨论过的缺陷——无论怎么也不能赚到消费者愿意支付的全部金额。例如，把价格定到 p，则消费者只愿意玩 q 次过山车（再玩的话还能获得的收益就不抵还需支付的费用了），他需要支付的就是图 7 中面积 $A+C$ 所代表的金额。而面积 C 是运行 q 次过山车的成本，因此商家最终只能赚到一个面积 A 的钱。而事实上，为了坐这 q 次过山车，消费者愿意支付的价格是面积 $A+C$ 再加上 A 上方的一个小三角形 F，那块面积 F 怎么能白白便宜了消费者呢？于是，商家想到，何不把那块小三角形面积以“门票”的形式一次性收入囊中呢？



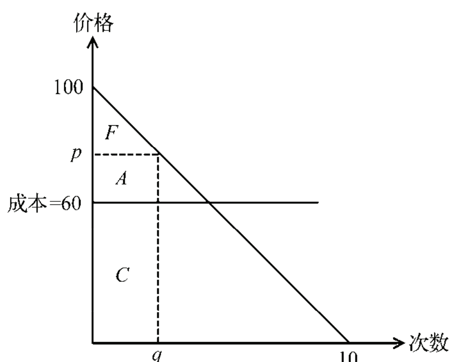


图 7

于是，我们有了收费的新方法：坐一次过山车的单价仍然是 p，但不管你坐多少次，你都需要事先缴纳面积 F 那么多钱作为门票。这样，你总共支付的价格就是面积 $A+C+F$，除去成本 C 后，商家赚到的部分就是面积 $A+F$。这就比刚才的定价方法多赚了一个三角形 F 的面积。

然而，这种方法仍然不是最好的。为了继续赚到区域 A 右边的那块面积，商家还可以降低过山车单价，让消费者再多坐几次过山车。最佳的两部分定价方案就是，把过山车的单价定得和成本一样，然后直接收取图 8 中成本线以上的整个大三角形面积 F' 的门票费。这样，消费者愿意坐 q' 次过山车，总共支付 $F'+C'$ 的钱，除去成本后商家净赚 F'，理论上把消费者榨取得一干二净。

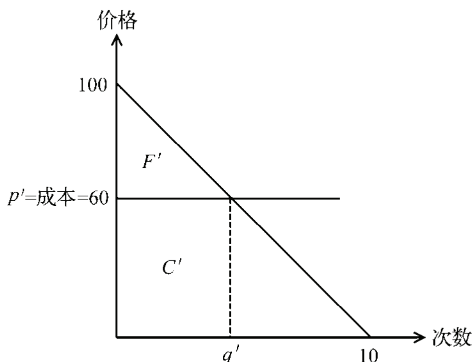


图 8



细心观察你会发现，在生活中，两部分定价的例子还有很多。会员入会费、信用卡年费、手机月租费都属于两部分定价的典型例子。

另一个常见的定价技巧叫做捆绑销售。例如，购买电视频道时，你会发现很多电视频道都不单卖，你必须要和其他的频道一起买才行。这就有些奇怪了：为什么不简单地按需求给每个频道订个价，而偏要费尽周折设计那么多频道包呢？难道打包卖会赚得更多一些吗？事实上还真是这样——捆绑销售会使得商家获得更高的利润。不论商家的行为如何诡异，其动机都是唯一的——赚尽可能多的钱。

为了解释这一现象，我们不妨从最简单的情况说起。假设有甲和乙两个人，以及A和B两个频道。甲愿意以120元购买A频道，愿意以30元购买B频道；乙只愿意以100元购买A频道，却愿意用40元购买B频道。如果对A和B两个频道分别定价，则显然应该为A频道定价100元，给B频道定价30元，此时商家收入260元。但若把A和B两个频道捆绑在一起销售，则可以给这个包定价140元，这能让商家收入280元。可见，捆绑销售确实能够给商家带来更多的利润。

但捆绑销售不见得总有效。如果把上面的数据稍稍更改一下，甲对两个频道的估值分别为120和40，乙对两个频道的估值分别为100和30，则单独定价和捆绑销售都只能收入260元，这之间并无差异。由此可见，不是随便两样东西捆绑起来就能带给商家更多利润的，这背后还隐藏有一些条件。

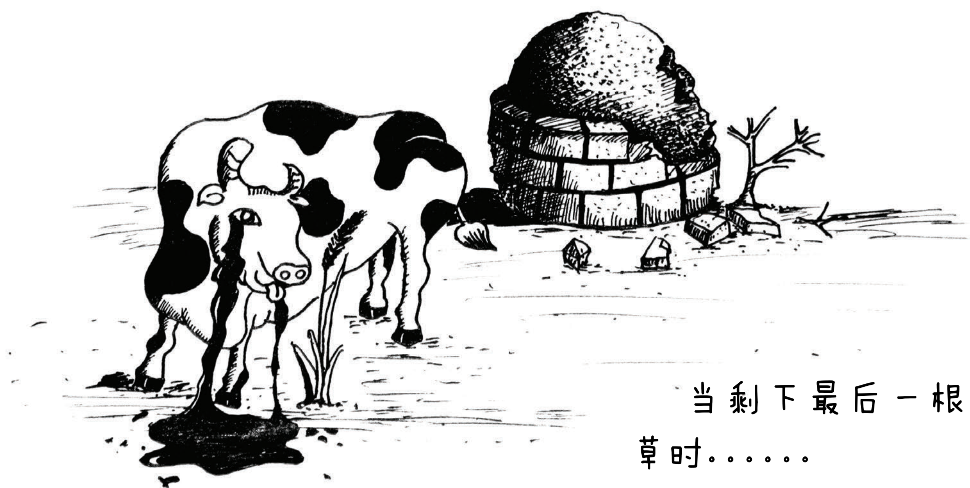
仔细观察你会发现核心问题所在——若捆绑销售能让商家赚更多，则一定是出现了这样的情况：这些频道的最低估价来自不同的买家，即买家对频道的评价不能是“都很好”或者“都不好”，对两个频道的评价呈现负相关。换句话说，对于某一系列商品，若消费者往往只偏爱其中一个，并且不同人的偏爱不同，则捆绑销售可以带来更多的利润。最经典的例子就是微软办公套件——为什么要把Word、Excel、PowerPoint捆绑销售，而不单卖呢？原因就在于，一个普通消费者并不会用到里面所有的软件，不同人对这几款软件的评价不同。虽然很多人觉得Word是最常用的，但财务人员会觉得Excel更加有用，而教师则觉得PowerPoint的价值更高。在这种情形下，捆绑销售将让商家赚更多的钱。重庆数字电视的特选节目包包含FOXTV、世界地理、发现之旅、第一剧场、风云音乐、英语辅导、风云足球和老故事这8个频道，频道内容覆盖面很宽，基本上满足上述条件。影剧院和游乐园的套票，颜色和款式不同但不单卖



的成套商品，都是典型的捆绑销售。

细心观察，你会发现生活中还有很多令人匪夷所思的定价策略，其实它们都是有企图有预谋的。维基百科上的“pricing strategies”条目中列举了 20 多种定价策略。如果每一个都用微观经济学来解释一番的话，估计又能写成一本书了。不过，上面这些例子就已经足以印证那句老话了：买的不如卖的精。

8. 公用品的悲剧



公用品悲剧是微观经济学中又一个非常有趣的话题。从一些简单的假设出发，通过一系列数学推导，我们能够得出一些乍看之下很不可思议的结论。利用这个结论，生活中很多反常的现象都有了合理的解释。

一个经典的公用品悲剧实例就是过度放牧的问题。同样一块牧场，如果为私人所有，牧场主将会非常合理地规划牧场，让放牧数量达到一个理论上的最优值；但是，如果这是一块公共牧场，则所有人都会争抢放牧，从而导致放牧数量远远大于最优值，最终每个人都得不到什么好处。可能有人会觉得这个现象并不难理解——既然是一块无人管制任人使用的公共牧场，人人都能在这里放牧，过度放牧自然就会不可避免地出现了。但是，仔细一想你会发现这个解释是有问题的：每一个来牧场放牧的人，自己心里也都知道，过度放牧对整个大局是不利的，自己的收益也会随之降低。既然人



人都知道过度放牧不好，为什么最后来放牧的人还是越来越多呢？私有牧场和公共牧场的区别到底在哪儿？我们可以借助数学工具来分析这个问题。

为了用数字来说明这一情况，我们首先做一些假设。我们假定牧场只放奶牛，收益也全部来自于牛奶供应。显然，牧场的总收益与放牧数量之间的关系是一个单峰函数——牧场上没有牛时总收益为 0，牛的数量超过牧场的最大容量后总收益也为 0，在这之间一定存在一个平衡点使得总收益达到最大。为此，我们无妨假设总收益 y 与放牛数量 x 满足 $y = x(100 - x)$ 的关系，即当牧场上的牛数为 0 或者为 100 时整个牧场都不会有任何收益，而 $x = 50$ 时牧场的总收益将会达到最大。我们再假设，购买一头牛的成本为 c，拥有奶牛之后放牧的成本则忽略不计。接下来，我们将求出该牧场在公有和私有两种情况下最终达到的放牧数量，大家将会看到开放牧场后确实将导致放牧数量远远超过最佳水平。

如果这是一块私有牧场，牧场主会选择放多少头牛呢？很多人可能会脱口而出，当然是 50 头牛，因为 $x = 50$ 时收益达到最大值。但请注意，牧场主想要最大化的并不是他的收入，而是减去成本后所得的利润 $x(100 - x) - cx$。对这个式子求导，就能得到利润最大化的条件：$100 - 2x - c = 0$。解出这个式子中的 x，就得到牧场中的最佳牛数 $x = \dfrac{100 - c}{2}$。

从另一个角度来看，上述结论也是很显然的：$100 - 2x$ 恰好就是 $x(100 - x)$ 的导数，是增加第 x 头牛给人带来的收入增加量。如果这个增加量比 c 大，那么买入一头新的牛显然划算；什么时候这个增加量比 c 小了，再买牛来放就要亏本了。因此，临界点 $100 - 2x = c$ 正好就是牛的数量达到最优的时候。

但是，一旦整个牧场变为公有，上述推理就不对了，因为单个放牧人并不关心整个牧场的利润，只在乎自己的盈亏。为了简便起见，我们假设牧场上有 x 个放牧人，每个人都只放 1 头牛。那么，牧场的总收入将为 $x(100 - x)$，每个人得到的收入为 $\dfrac{x(100 - x)}{x} = 100 - x$。因此，当牧场上有 $x - 1$ 头牛时，对这块蛋糕垂涎已久的人会发现，他作为第 x 个放牧人进入牧场后，能够分得的收入为 $100 - x$，只要这个值比 c 大，这样做就是值得的。随着进入牧场的人数增多，新加入的放牧人会发现他所能赚到的越来越少。最终当 $100 - x = c$ 时，便不会再有人想要进入该牧场了。此时的总体情况

惨不忍睹——每个放牧人所得的收入都是 c ，可以说是一分钱也赚不到。

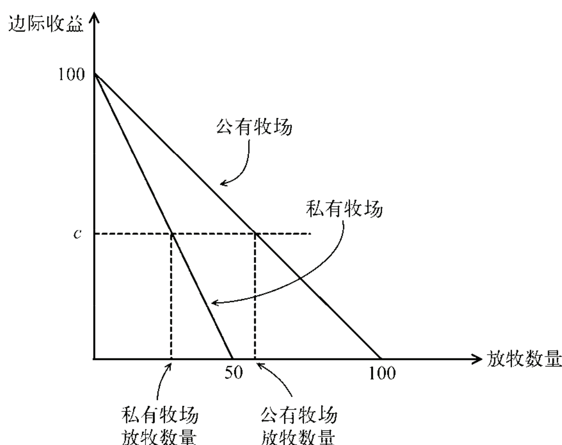


图　1

为什么同样都是为了自己的利益最大化，公有牧场和私有牧场的差别那么大呢？根本原因就在于，在公有牧场中，每个人都能自由地进出该牧场，每个人都拥有在牧场放牧的权利。只要一个新来的放牧人发现自己有钱赚，他就会选择买牛放牧，而并不关心这样做其实会导致每个已经在牧场上的放牧人都要少赚一些。但是，选择在这里放牧是这个放牧人的权利，原来的放牧人没有理由驱逐他。随着新人不断加入，每个人都会赚得越来越少，最后大家的利润就将趋于 0，悲剧也就产生了。

不仅仅是公共牧场，事实上“公用品悲剧”发生在几乎所有的公共资源上。例如，人人都知道污染环境损人损己，最后弄得每个人都活不下去，但为什么大家仍然亡了命似地破坏自然资源呢？原因就在于，对于某一个企业来说，直接排放污水废气给它带来了一个正的收益，但这却给其他的每个企业都造成了损失。每个决定要污染环境的企业都这么想，这样做的人便会越来越多，整个社会的损失也就越来越大。

公用品悲剧还会涉及很多非自然资源的公用品。例如，每个人都知道公交车挤着不舒服，但为什么最终车上还是这么挤呢？这就是因为，对于每个车下面的人来说，只要能上车，他就已经得到了好处，完全无视这一举措会使车上的每个人都受到一点损失。每个车下的人都这样想，悲剧也就发生了。

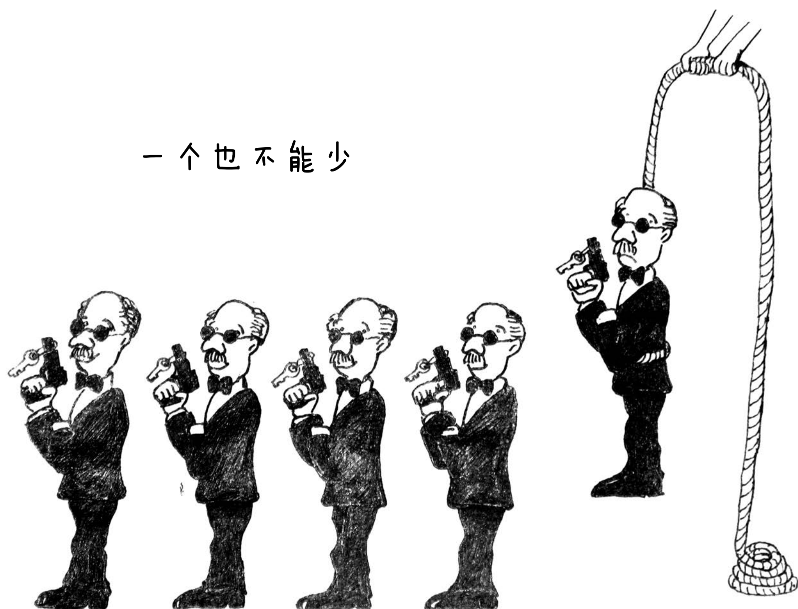
公用品悲剧的理论还有很多更奇怪的应用。很多时候，交通堵塞的原因是前方路



段发生车祸，但事实上前方发生的仅仅是某辆车撞上了护栏，车祸根本没有挡住道路。为什么最终还是堵车了呢？原来，每辆开到车祸现场的车，都会减慢速度看看热闹，甚至停下来掏出手机照下这一“杰作”。这样做虽然满足了自己的好奇心，却让后面的每一辆车都多堵上好几秒。因此大家往往会发现这样一个有趣的现象：车祸越离奇，交通堵塞越厉害。另一个经典而有趣的应用是，为什么在餐桌上，实行 AA 制的总消费要比某一个人请客的总消费高出许多。原因就在于，在实行 AA 制后，每个点菜的人都会想，原来需要一百多块钱才能吃到的美味，这次只需要花二三十块钱便能享受到了。这样，虽然自己得到了满足，却让每个人都为你多付了一些钱。



9. 密码学与协议



说到“密码学”，大多数人的第一念头或许是摩尔斯（Morse）电码、凯撒（Caesar）移位密码以及同音替换密码之类的东西。这些东西在各类小说中都已是老面孔了，“字母 e 在英文中出现的频率最高”等基本的破密码方法已经是耳熟能详了。某次和网友云风[①]聊了一下，突然领会到了密码学的真谛。密码学关注更多的并不是加密解密的各种数学算法，而是在已有数学算法上如何实现各种安全需求。防止消息泄露只是众多安全问题中的冰山一角，而这个问题本身又有很多复杂的变化。

谈到“消息泄露”时，我们头脑中想到的往往是，在信息传输过程中如何防止第

① 云风的博客地址为 http://codingnow.com。



三方截获。当然，小偷防是防不住的，不过我能保证他偷到东西也没用。双方只需要事先约定一套加密解密的方法，以密文的方式进行传输，这样便能很好地防止消息泄露。但有时候，“消息泄露”的内涵复杂得多，加密解密的传统方法并不适用。考虑这么一个问题：10 个人坐在一起谈天，突然他们想知道他们的平均年薪是多少，但每个人都不愿意透露自己的工资数额。有没有什么办法让他们能够得出答案，并且不用担心自己的年薪被曝光？事实上，最简单的解决办法不需要依赖任何密码学知识：第一个人随便想一个大数，比如 880 516。然后他在纸条上写下这个数与自己的年薪之和，传给第二个人；第二个人再在这个数上加上他的年薪数额，写在另一张纸条上传给第三个人；直到最后一个人把纸条传回第一个人后，第一个人把纸条上的数减去只有自己知道的那个 880 516，便得到了全部 10 个人的工资和。

可以看到，密码学不仅仅研究加密解密的数学算法。更多的时候，密码学研究保护信息安全的策略，我们可以称之为“协议”。在已有的数学模型基础上，我们往往忽略具体的数学实现方法，转而专注地研究借助这些数学工具能够构建的安全措施。除了消息保密性以外，密码学还研究一些更加有趣的问题。这里，就让我们一起来看四个有意思的密码学协议问题吧。

首先我们来看一个日常生活中大家经常会遇到的密码学协议问题——签合同。签署合同会具备法律效应，人们往往不敢随意签名。合同一般同时规定了双方的权利和义务，并需要双方都在上面签名。第一个在上面签字的人就会觉得很亏：万一我签了字后对方突然翻脸耍赖不签了咋办？即使合同上规定“合同仅在双方均签署之后才有效”，这个问题仍然存在，因为后签名者将具有绝对的主动权，他想什么时候签就什么时候签，而只有他的签名才具有决定意义。因此很多时候，双方都希望能够在对方签名之后自己再签名，从而获得一些安全感。这里我们来探讨一个有趣的问题：有没有什么办法能够让双方同时签约，使得双方签名时都能确保自己的利益安全？

如果我们谈论的是传统意义上的签名，同时签名当然是有可能办到的：双方只需要拿起各自的笔，同时在文件上写下自己的名字即可。当然，事实上肯定不会有人这么做。试想这样一个荒唐的画面：两个西装笔挺的人挤在一起，两只手臂磕磕碰碰地交错在一起，然后双方同时喊“三、二、一”并一起开始写字……比起自己丢掉的脸面，自己先签名所带来的忧虑似乎也不算什么了。

有没有体面一些的，不那么荒唐的同时签字法呢？这里有一个很有启发性的办法。合同双方面对面地坐在桌子的两端。其中一个人在合同上写下自己名字的第一个字母，然后传给坐在对面的第二个人；第二个人写下他自己名字的第一个字母，然后又递回给第一个人；第一个人签下自己名字的第二个字母，再交给对方要求他写下他的名字的第二个字母……以此类推，直到双方都签署完自己的名字为止。为了让双方能够“同时”签完，名字较长的人偶尔可能需要连续写下两三个字母。换成汉字的话，这个方法同样适用——把字母改成笔划就行了。

双方都愿意履行这一协议，因为在这个协议下双方是一点一点地签完整个文件的。第一个写字的人不会觉得自己很亏，因为写下一个字母是远不具备法律效应的；如果对方拒绝签他的第一个字母，我可以当即撕毁合同。虽然他们都不知道，究竟要写多少个字母才算签字，但大家都保持自己签下的名字长度与对方基本相当，因此不会担心对方突然放弃协议。就在这种互动的心理过程中，签名的法律效应一点一点地增强，直到最后双方写完自己的名字。

但是，这个办法不能用于数字签名。利用电子加密算法进行签名是一个整体的过程，不能一部分一部分地进行。能不能把合同拆成若干份，然后双方一份一份地逐个签名呢？当然不行。如果某一份合同里有一个至关重要的义务性条款，后签名的人等对方签到这里后便可以立即终止签名，从而谋取利益。那么，能不能规定“仅当你把所有 n 个部分的文件都签过了才算签”呢？这意味着最后一次签名才具有最终的决定意义，说穿了不过是把安全问题转移到了“谁签最后这一下”，问题实质上并未改变。其实，我们的解决办法相当简单。我们可以要一个小花招，从本质上模拟上面的“逐字母签名法”。

首先，第一个人签署这样一份文件：“我愿意以 1%的概率接受该合同。”第二个人检查第一个人的签名，然后在上面附加一句“我愿意以 2%的概率接受这份合同”，并进行签名，再交回给第一个人。第一个人检查对方已经签名，然后继续将这个条文升级为“我愿意以 3%的概率签署该合同”并签名。双方来来回回签 100 次，直到最后第一个人签“我愿意以 99%的概率签署这份合同”，然后轮到第二个人签署“我接受该合同”，最后再轮到第一个人签署“我接受该合同”。



注意，这个“接受概率”是有实际意义的。如果在第一个人第一次签完文件后，第二个人立即放弃继续签署，法官可能会要求双方进行一次公开抽签测试，选取一个不超过 100 的正整数。如果这个数恰好为 1，那么签署这句话的人就相当于签署了这份合同。类似地，我们也可以约定，当一人声称将以百分之$(p-1)$的概率接受此合同，另一人声明以百分之 p 的概率接受时，法官可以要求双方共同生成一个在 1 和 100 之间的整数：如果它不超过$(p-1)$则双方都接受，如果它的值比 p 大则双方都不接受，若它的值正好等于 p 则合同仅被后者接受。因此，这种协议实质上是用概率法再现了“逐字母签名法”的核心思想，将法律效应的问题进行量化，使得率先签名的潜在危险减小到了原来的百分之一。

合同签署的问题就说到这里了。让我们再来看另外一个有趣的协议问题。设想这样一个场景：总部打算把一份秘密文件发送给 5 名特工，但直接把文件原封不动地发给每个人，很难保障安全性。万一有特工背叛或者被捕，把秘密泄露给了敌人怎么办？于是就有了电影和小说中经常出现的情节：把绝密文件拆成 5 份，5 名特工各自只持有文件的$\frac{1}{5}$。不过，原来的问题并没有彻底解决，我们只能祈祷坏人窃取到的并不是最关键的文件片段。因此，更好的做法是对原文件进行加密，每名特工只持有密码的$\frac{1}{5}$，这 5 名特工需要同时在场才能获取文件全文。但这也有一个隐患：如果真的有特工被抓了，当坏人们发现只拿到其中一份密码没有任何用处的同时，特工们也会因为少一份密码无法解开全文而烦恼。此时，你或许会想，是否有什么办法能够让特工们仍然可以恢复原文，即使一部分特工被抓住了？换句话说，有没有什么密文发布方式，使得只要 5 个人中半数以上的人在场就可以解开绝密文件？这样的话，坏人必须要能操纵半数以上的特工才可能对秘密文件造成实质性的影响。这种秘密共享方式被称为 $(3,5)$ 门限方案，意即 5 个人中至少 3 人在场才能解开密文。

实现 (m,n) 门限方案的一个传统办法是，把这份文件的密码拆成 C_n^{m-1} 份，每个人持有 C_{n-1}^{m-1} 份密码。不妨假设文件的密码是一个 100 位数，那么在 $(3,5)$ 门限方案中，我们需要把这个密码拆成 $C_5^2=10$ 份，每份密码都是一个 10 位数。不妨把这 10 份密码分别用 0 到 9 编号，把每份密码都额外复制两份。5 名特工各持有 6 份密码，密码的分配如下：



特工#1	0	1	2	3	4	5				
特工#2	0	1	2				6	7	8	
特工#3	0			3	4		6	7		9
特工#4		1		3		5	6		8	9
特工#5			2		4	5		7	8	9

你可以自己验证一下：任意 3 名特工碰头，都能凑齐这 10 份密码；但任意 2 名特工碰头，都无法凑齐所有的密码。

上述分配表的构造其实很简单：给每个可能的“三人组”分配一份密码。从 5 个特工中选出 3 个人共有 10 种方案，因此我们正好要 10 份密码。例如把密码 0 分给特工 1、2、3，把密码 1 分给特工 1、2、4，一直到把密码 9 分给特工 3、4、5。这样的话，任意 2 个人在场都无法打开文件，因为他们始终缺少一份密码（这份密码分给了其余 3 个人）。而任意 3 个人在场都足以打开文件，因为每一份密码都只缺少 2 个人的量，不可能出现这 3 个人都没有的情况。这样，我们便利用组合数学巧妙地解决了这一问题。

在密码学中，我们有一些更精妙的方案。最巧妙的方法是，把文件密码编码为三维空间中的一个点，然后生成 5 个过该点的平面，每个特工持有其中一个平面方程。显然，2 个特工在一起是无法获得原文件的，因为 2 个平面的公共点有无穷多个；但 3 个平面的交点是唯一的，因此任意 3 个人在一起都能解开原文件。

另一个有趣的办法利用了下面这个事实：知道 $m-1$ 次多项式函数上的任意 m 个点就能恢复出整个多项式。因此，我们可以把文件密码编码为一个二次多项式 $f(x)$，然后把 $f(1)$、$f(2)$、$f(3)$、$f(4)$ 和 $f(5)$ 的值告诉对应的特工。任意 3 个特工碰头之后，只需要解出这个多项式 $f(x)$ 即可恢复出文件的密码来。

上述两种方案的本质都是相同的：把文件密码设为 3 个数 x、y、z，然后编写 5 个与 x、y、z 有关的一次方程，并把这 5 个方程分别交给 5 名特工。假如文件的密码是 116.35、39.975、67.167 这 3 个数，只有同时输入这 3 个数，才能解开原文件。那么，我们就用这 3 个数编写五个三元一次方程：

$$3.4x+5.6y-2.81z=430.711$$

$$x-2.11y+0.09z=38.0478$$



$$7x+9.9y-0.1z=1203.49$$
$$-0.3x+2.24y+5.6z=430.774$$
$$3x+4.5y+6.67z=976.941$$

其中 $x=116.35$ 、 $y=39.975$ 、 $z=67.167$ 是它们的公共解。但是，要想确定出这个公共解，只有 1 个方程或者 2 个方程是不够的。事实上，至少需要 3 个方程，才能保证三元一次方程组存在唯一解。因此，至少需要 3 个人在场，才能获得秘密文件的密码。

利用数论知识我们还能得到一个简单的协议。中国剩余定理告诉我们，给出 m 个两两互质的整数，它们的乘积为 P ；假设有一个大整数 M ，如果我们已知 M 分别除以这 m 个数所得的余数，那么在 0 到 $P-1$ 的范围内可以唯一地确定这个 M 。我们可以想办法构造这样一种情况，n 个数之中任意 m 个的乘积都比 M 大，但是任意 $m-1$ 个数的乘积就比 M 小。这样，任意 m 个除数就能唯一地确定 M ，但 $m-1$ 个数就不足以求出 M 来。米尼奥（Mignotte）门限方案就用到了这样一个思路。我们选取 n 个两两互质的数，使得最小的 m 个数的乘积比最大的 $m-1$ 个数的乘积还大。例如，在 (3,5) 门限方案中，我们可以取 53、59、64、67、71 这 5 个数，前面 3 个数乘起来得 200 128，而后面两个数相乘才得 4757。我们把文件的密码设为一个 4757 和 200 128 之间的整数，比如 123 456。分别算出 123 456 除以上面那 5 个数的余数，得到 19、28、0、42、58。显然，知道任意 3 个同余方程就可以唯一地确定出 123 456，但仅知道 2 个方程只能得到成百上千个不定解。例如，假设我们知道了 x 除以 59 余 28，也知道了 x 除以 67 余 42，那么我们只能确定在 0 和 $59\times67-1$ 之间的解 913，并且只能断定 M 是一个形如 $59\times67\times k+913$ 的数，其中 k 的数量级和当初选的那五个数一样大。

我们的第三个协议问题就更有意思了。3 个好朋友到一家餐厅吃饭。饭快吃完的时候，一个服务员过来告诉他们说，他们的账单已被匿名支付了。3 个人都尊重他人匿名付款的权利，但同时他们也想知道，这个匿名支付者是他们三位中的一个，还是他们三人之外的某个第四者。有没有什么办法能够让他们知道在他们中间是否有人付账，但又保证任何人都推测不出究竟是谁付的账？利用 3 枚硬币就能轻易做到这一点。

假设这 3 个人围着一张圆桌坐成一圈。每个人都在自己和右手边那个人中间抛掷



一枚硬币，并用另一只手挡住硬币，使得这枚硬币只有他俩才看得见。这样的话，每个人都只能看见他左右的两枚硬币（但看不见桌子对面的第三枚硬币）。每个人都大声报出，自己身边的两枚硬币的正反面是否相同。如果他们中间有人付账，则这个人报出与实际情况相反的词，相同的话说“不同”，不同的话则说“相同”。显然，如果大家说的都是真话，则报“不同”的次数一定是偶数次。如果有奇数个人说“不同”，那么一定有一个人说假话，这表明匿名支付账单的人就在他们中间。

注意到这个方案可以扩展到 n 个人。我们只需要证明，假如有 n 个人坐成一圈，如果大家都说真话，则说“不同”的次数一定是偶数次。证明非常简单。想象你从某一枚硬币出发，顺时针查看每一枚硬币的正反，得到一个硬币正反序列。每当这个序列由正变反或者由反变正时，就相当于有一处“不同”的情况发生。然而，当你绕着圆桌走完一圈，回到出发点时，硬币序列又变回了出发时的正反。因此，途中发生的“不同”次数一定是偶数次。

其实，抛掷硬币只是一个形象的描述方法罢了。在没有硬币，甚至大家根本没坐在一起的情况下，这个协议也很容易实施。比方说，先在网上公布整个协议规则，并约定一个虚拟的座位顺序；然后每个人都在 1 和 2 之间想一个数，并把结果以短信的形式发给他右边的那个人；最后每个人都按照协议规则，在网上发一个“相同”或者“不同”。

这个协议有一个意想不到的用途——匿名的消息广播。假如一群人围坐成一圈开会，会议过程中需要在场的一个不愿透露自己身份的人进行匿名发言。为此，大家可以统一采用上面的抛硬币协议（或者对应的电子协议，只是为了简便，下面还是采用抛硬币的说法）。发言人将信息编码为一个长度为 n 的 01 串。硬币投掷分 n 轮进行。第 i 轮中，其他人都严格按照实际情况报是否相同，发言人则根据编码信息的第 i 位的值来通报：若第 i 位为 0，则按照实际情况通报；若第 i 位为 1，则说与实际情况相反的词。这样，实际的信息就应该是每轮叫“不同”的次数除以 2 的余数形成的序列。

我们把最有意思的话题放在了最后。现在，假如你碰到了一个宣称可以预知未来的人，他说他知道下周的彩票中奖号码。你肯定不会相信，便用激将法让他说出下周的中奖号码：“你说出来啊，你要是说不出来，那就表明你不能预测未来。”不过，他却一本正经地说：“不行，我虽然能预测未来，但却不能把它说出来，否则会产生蝴



蝶效应，改变这个宇宙既定的将来，导致危险的时空悖论。”

哈哈，这个“先知”真是天才呀！能预言未来却不能说出来，这样就永远不能证伪了。

不过，治他的方法也不是没有。比方说，可以叫他把预测结果写在一张纸上，锁进一个盒子里。然后，你拿走盒子，他拿走钥匙。彩票中奖号码公布后，你们再碰个头，把盒子打开，来看看当初的预测结果是否正确。这样便能让他做出一个谁都不能看见，但他今后也不能抵赖的预测。我们把这样的协议叫做“带有防欺骗的承诺”。

只可惜，这种方法有一个局限性：它只能在现实生活中使用。如果你在网上遇到了自称能预知未来的人，你如何让他做出防欺骗的承诺呢？

有人可能会说，为什么不让他给预言加一个密呢？就像之前让他给预言加上一把锁一样。比方说，让他在下周的中奖号码上加一个很大的整数，然后把结果告诉你；这个很大的整数就是解开中奖号码的密钥，由他自己保管。仔细想想你会发现，这个方案显然不行，因为到了验证预言的时候，他可以谎报这个大整数，让密码解开来后是任何一个他想要的数。为了防止他耍赖，能否让他事先就把密钥公布出来呢？这也不行——知道了密钥后，你就能直接获得密码原文了，这样便失去了保密的作用。

注意到，传统的加密方法不能公开的原因就是，知道了加密方法也就知道了解密方法，只需要把加密方法反过来做就行了。有没有一种加密方法，使得即使你知道了加密的方法，也不能恢复出密码原文呢？有的。只需要在加密过程中加入一些不可逆的数学运算就行了。比方说，你们可以约定这样一种加密方法：先取中奖号码的正弦值的小数点后八位数字，得到一个八位整数；再求中奖号码与圆周率前六位数字形成的整数（314 159）之和，取该和的平方的第 3 位到第 10 位，又得到一个八位数；最后计算这两个八位数的和除以 456 789 的余数。假如他预言的中奖号码是 1 234 567，那么对 1 234 567 进行上面这一串操作后，将会得到 244 685。但是，即使知道加密的过程，你也不能把 244 685 还原成 1 234 567。事实上，1 234 567 甚至不一定是唯一解，很可能有别的数加密后也会变成 244 685。上述加密方法能把任何数都加密成一个小于 456 789 的数，因此必然会出现不同的数加密成同一结果的情况。这就意味着，这种加密方法是会丢掉原始信息的。我们不妨把这种不可逆的加密方法叫做“单向加密”。在密码学中，MD5 和 SHA1 是两种比较常用的单向加密算法。由于其单向性，

这种加密方法不能用于普通的信息传输。但它有很多其他的应用，做出带有防欺骗的承诺便是一例。拿到 244 685 这个数后，你完全无法推出他究竟做了什么预测；到了验证预言的时候，只需要让对方宣布当初他的预测 1 234 567，你来检验一下 1 234 567 加密后是否会得到 244 685 就行了。

不过，这个方法有一个局限性：如果他宣称他能预测某只股票会涨还是会跌，上述方法就有漏洞了。比方说，你们可以约定，数字 1 表示股票会涨，数字 2 表示股票会跌，然后让他用刚才的那套方法把他的预测结果加密发过来。如果他告诉你的结果是 316 554，那你只需要分别试一下 1 和 2 加密后分别得多少，就知道原始数据是 1 还是 2 了。原始数据的取值太有限，让穷举式的“暴力破解”变得易如反掌。怎么办呢？可以想办法硬把原始数据的取值范围扩大。比如，约定所有个位数字为 1 的数都表示股票会涨，约定所有个位数字为 2 的数都表示股票会跌。假如对方预测股票会涨，他可以选取任意一个末位为 1 的数，对其进行加密，这下你便没办法暴力破解了。不过，这里还有一个小问题：刚才我们说了，单向加密可能会把不同的原始信息加密成同一个结果，因此完全有可能出现这样两个数，它们的末位分别是 1 和 2，但加密后的结果相同（虽然找到这样的例子并不容易）。为了避免对方手中持有精心构造的“两可解”情况，我们可以在每次实施协议时都改变一下协议的细节，比如每次都换一种单向加密方式，或者更好地，每次都要求对方选取的那个数必须以你想的某个随机数打头。这样一来，整个协议就完美了。

其实，这个协议并不只在揭穿超能力者的时候才有用。我们生活当中有很多地方都可以用到带有防欺骗的承诺。有一次，我在战网上和别人打星际，打出了一个非常搞笑的局面：两边的兵都一个不剩，两边的钱也都不够造东西了，双方都完全丧失了战斗能力。但是，双方都还剩有建筑，因此都不算输。此时，必须有一个玩家主动认输，先退出游戏，才能结束僵局。该谁先退呢？我和他便在游戏中互发消息谈论了起来。其实，在现实生活中，这很好解决：来玩一次石头剪子布就可以了。但是，怎么在网上玩石头剪子布呢？总不能让一个人先发消息说“我出的是剪子”，另一个人回复“哈哈，我出的石头”吧。这时候，就要用到带有防欺骗的承诺了。我们可以利用前面讲的方法，各自向对方承诺自己要出什么拳，然后双方再公布自己出的拳，让对方验证自己并没撒谎。更简单的方法就是，我在 1 和 2 之间想一个数，然后把我想的



数加密告诉对方，由对方来猜我想的数是多少，猜对了我就认输退出，猜错了他就认输退出。对方做出猜测后，我再公布加密前的原始信息，以证明我没有耍赖。

我们常常在电视上看到这样的一幕：一位老太太兴冲冲地走上台去，翻过 3 个商标牌，发现上面尽是 5 块钱、10 块钱的小奖，垂头丧气地回到观众席；然后主持人会跑出来，边翻着另外几个牌子边说，1000 元的大奖在这个后面，800 元的在这里之类的。为什么主持人要演出“事后揭大奖”这一幕呢？道理很简单，节目组想通过这个“验证过程”告诉观众，这个环节不是骗人的，大奖真的就在这里面，只是刚才那家伙运气背了没摸到而已。摸奖前宣称有大奖，摸完奖之后还能证实大奖真的存在，这也是带有防欺骗的承诺。

但是，我们喝饮料参与开盖有奖活动时，就会有被欺骗的感觉：你说中奖率是千分之一，我凭什么相信你呢？那么，有没有办法让开盖有奖活动的中奖率变得透明呢？有的。我就想过这么一个方法。比如说，开盖后你将得到一个参与活动的序列号，把这个序列号短信发送给活动举办方参与抽奖。此时，活动举办方的服务器从 1 到 1000 中随机生成一个整数，并把这个整数加上你指定的前缀和它自选的前缀，用公开的单向加密方法加密后发回给你。你需要猜出服务器生成的数是什么，如果猜对就能中奖，如果猜错就结束游戏。发送了你的猜测结果后，服务器将发来加密前的信息，确保自己没有撒谎。

密码学与协议的故事多得讲也讲不完。公钥加密算法、密钥交换协议、盲签名协议、投票协议、虚拟货币协议、中间人攻击……这些简直都是密码学中的珍宝。还没过瘾的读者，不妨买一本密码学与协议的书，继续研究下去。

10. 公平分割问题

大家或许都知道经典的两人分饼问题——为了实现公平性，只需要一个人切，另一个人选即可。不过，在现实生活中，情况远没有那么理想。如果把大饼换成蛋糕，问题就复杂了很多——你想吃奶油，我想吃巧克力，他想吃水果——如果分蛋糕的人对蛋糕各部分的价值看法有分歧，还能实现公平的分割吗？如果分蛋糕的人不止两个呢？

事实上，对于两个人分蛋糕的情况，经典的“你来分我来选”的方法仍然是非常有效的，即使双方对蛋糕价值的计算方法不一致也没关系。首先，由其中一人执刀，把蛋糕切分成两块；然后，另一个人选出他自己更想要的那块，剩下的那块就留给第一个人。由于分蛋糕的人事先不知道选蛋糕的人会选择哪一块，为了保证自己的利益，他必须（按照自己的标准）把蛋糕分成均等的两块。这样，不管对方选择了哪一块，他都能保证自己总可以得到蛋糕总价值的$\frac{1}{2}$。

不过，细究起来，这种方法也不是完全公平的。对于分蛋糕的人来说，两块蛋糕的价值均等，但对于选蛋糕的人来说，两块蛋糕的价值差异可能很大。因此，选蛋糕的人往往能获得大于$\frac{1}{2}$的价值。一个简单的例子就是，蛋糕表面是一半草莓一半巧克力的。分蛋糕的人只对蛋糕体积感兴趣，于是把草莓的部分分成一块，把巧克力的部分分成一块；但他不知道，选蛋糕的人更偏爱巧克力一些。因此，选蛋糕的人可以得到的价值超过蛋糕总价值的一半，而分蛋糕的人只能恰好获得一半的价值。而事实上，更公平一些的做法是，分蛋糕的人得到所有草莓部分和一小块巧克力部分，选蛋糕的人则分得剩下的巧克力部分。这样便能确保两个人都可以得到一半多一点的价值。

但是，要想实现上面所说的理想分割，双方需要完全公开自己的信息，并且要能



够充分信任对方。然而，在现实生活中，这是很难做到的。考虑到分蛋糕的双方尔虞我诈的可能性，实现绝对公平几乎是不可能完成的任务。因此，我们只能退而求其次，给“公平”下一个大家普遍能接受的定义。在公平分割（fair division）问题中，有一个最为根本的公平原则叫做“均衡分割”（proportional division）。它的意思就是，如果有 n 个人分蛋糕，则每个人都认为自己得到了整个蛋糕至少 $\frac{1}{n}$ 的价值。从这个角度来说，“你来分我来选”的方案是公平的——在信息不对称的场合中，获得总价值的一半已经是很让人满意的结果了。

如果分蛋糕的人更多，均衡分割同样能够实现，而且实现的方法不止一种。其中一种简单的方法就是，每个已经分到蛋糕的人都把自己手中的蛋糕分成更小的等份，让下一个没有分到蛋糕的人来挑选。具体地说，先让其中两个人用“你来分我来选”的方法，把蛋糕分成两块；然后，每个人都把自己手中的蛋糕分成三份，让第三个人从每个人手里各挑出一份来；然后，每个人都把自己手中的蛋糕分成四份，让第四个人从这三个人手中各挑选一份；不断这样继续下去，直到最后一个人选完自己的蛋糕。只要每个人在切蛋糕时能做到均分，无论哪块被挑走，他都不会吃亏；而第 n 个人拿到了前面每个人手中价值至少 $\frac{1}{n}$ 的小块，合起来自然也就不会少于蛋糕总价值的 $\frac{1}{n}$ 。虽然这样下来，蛋糕可能会被分得零零碎碎，但这能保证每个人手中的蛋糕在他自己看来都是不小于蛋糕总价值的 $\frac{1}{n}$ 的。

还有一种思路完全不同的分割方案叫做“最后削减人算法”，它也能做到均衡分割。我们还是把总的人数用字母 n 来表示。首先，第一个人从蛋糕中切出他所认为的 $\frac{1}{n}$ ，然后把这一小块传给第二个人。第二个人可以选择直接把这块蛋糕递交给第三个人，也可以选择从中切除一小块（如果在他看来这块蛋糕比 $\frac{1}{n}$ 大了），再交给第三个人。以此类推，每个人拿到蛋糕后都有一次“修剪”的机会，然后移交给下一个人。规定，最后一个对蛋糕大小进行改动的人将获得这块蛋糕，余下的 $n-1$ 个人则从头开始重复刚才的流程，分割剩下的蛋糕。每次走完一个流程，都会有一个人拿到了令他满意的蛋糕，下一次重复该流程的人数就会减少 1。不断这样做下去，直到每个人都



分到蛋糕为止。

第一轮流程结束后，拿到蛋糕的人可以保证手中的蛋糕是整个蛋糕价值的$\frac{1}{n}$。而对于每个没有拿到蛋糕的人来说，由于当他把蛋糕传下去之后，他后面的人只能减蛋糕不能加蛋糕，因此在他看来被拿走的那部分蛋糕一定不到$\frac{1}{n}$，剩余的蛋糕对他来说仍然是够分的。在接下来的流程中，类似的道理也同样成立。更为厉害的是，在此游戏规则下，大家会自觉地把手中的蛋糕修剪成自认为的$\frac{1}{n}$，要赖不会给他带来任何好处。分蛋糕的人绝不敢把蛋糕切得更小，否则得到这块蛋糕的人就有可能是他；而如果他把一块大于$\frac{1}{n}$的蛋糕拱手交给了别人，在他眼里看来，剩下的蛋糕就不够分了，他最终分到的很可能远不及$\frac{1}{n}$。

这样一来，均衡分割问题便完美解决了。不过，正如前面我们说过的，均衡条件仅仅是一个最低的要求。在生活中，人们对"公平"的概念还有很多更不易形式化的理解。如果对公平的要求稍加修改，上述方案的缺陷便暴露了出来。让我们来看这样一种情况：如果n个人分完蛋糕后，每个人都认为自己分得了至少$\frac{1}{n}$的蛋糕，但其中两个人还是打起来了，可能是什么原因呢？由于不同的人对蛋糕各部分价值的判断标准不同，因此完全有可能出现这样的情况——虽然自己已经分到了至少$\frac{1}{n}$份，但在他看来，有个人手里的蛋糕比他还多。看来，我们平常所说的公平，至少还有一层意思——每个人都认为别人的蛋糕没我手里的好。在公平分割理论中，我们把满足这个条件的分蛋糕方案叫做免嫉妒分割（envy-free division）。

免嫉妒分割是一个比均衡分割更强的要求。如果每个人的蛋糕都没我多，那我的蛋糕至少有$\frac{1}{n}$，也就是说满足免嫉妒条件的分割一定满足均衡条件。但反过来，满足均衡条件的分割却不一定是免嫉妒的。比方说，A、B、C 三人分蛋糕，但 A 只在乎蛋糕的体积，B 只关心蛋糕上的草莓颗数，C 只关心蛋糕上的巧克力块数。最后分得的结果是，A、B、C 三人的蛋糕体积相等，但 A 的蛋糕上什么都没有，B 的蛋糕上有一颗草莓两块巧克力，C 的蛋糕上有两颗草莓一块巧克力。因此，每个人从自己的角



度来看都获得了整个蛋糕恰好 $\frac{1}{3}$ 的价值，但这样的分法明显是不科学的——B、C 两人会互相嫉妒。

之前我们介绍的两种均衡分割方案，它们都不满足免嫉妒条件。就拿第一种方案来说吧，如果有三个人分蛋糕，按照规则，首先应该让第一人分第二人选，然后两人各自把自己的蛋糕切成三等份，让第三人从每个人手中各挑一份。这种分法能保证每个人获得至少 $\frac{1}{3}$ 的蛋糕，但却可能出现这样的情况：第三个人从第二个人手中挑选的部分，恰好是第一个人非常想要的。这样一来，第一个人就会觉得第三个人手里的蛋糕更好一些，这种分法就不和谐了。

构造一套免嫉妒的分割方案非常困难。1960 年，约翰·塞尔弗里奇（John Selfridge）和约翰·康威（John Conway）[①]各自独立地分析了人数为 3 的情况，构造出了第一个满足免嫉妒条件的三人分割方案。这种分割方案就被称为“塞尔弗里奇–康威算法”。

首先，A 把蛋糕分成三等份（当然是按照自己的看法来分的，后面提到的切分和选取也都是这样）。如果 B 认为这三块蛋糕中较大的两块是一样大的，那么按照 C、B、A 的顺序依次选取蛋糕，问题就解决了。麻烦就麻烦在 B 认为较大的两块蛋糕不一样大的情况。此时，B 就把最大的那块蛋糕的其中一小部分切下来，让剩余的部分和第二大的蛋糕一样大。被切除的部分暂时扔在一旁，在第二轮分割时再来处理。接下来，按照 C、B、A 的顺序依次选蛋糕，但有一个限制：如果 C 没有选那块被修剪过的蛋糕，B 就必须选它。

这样，三人就各分得了一块蛋糕。由于 A 是切蛋糕的人，对于他来说拿到哪一块都一样，因此 A 不会嫉妒别人。由于 B 选取的是两个较大块中的一个，因此 B 也不会嫉妒别人。由于 C 是第一个选蛋糕的，显然他也不会嫉妒别人。因此，就目前来说，三个人之间是不会有嫉妒发生的。

但是，还有一小块被切除的部分没分完，因此分割流程进入第二轮。

在 B 和 C 之间，一定有一个人选择了那块被修剪过的蛋糕。不妨把这个人重新记作 X，另一个人就记作 Y。让 Y 把最后那一小块分成三等份，按照 X、A、Y 的顺序

---

① 康威生于 1937 年，是一位非常有名的英国数学家。他发明并研究了很多有趣的数学游戏。记住康威这个名字，后面我们还会多次提到他。



依次挑选蛋糕，结束第二轮流程。这一轮结束后，每个人都又得到了一小块蛋糕。由于 X 是第一个选蛋糕的人，X 显然不会嫉妒别人；由于 Y 是分蛋糕的人，Y 也不会嫉妒别人。由于 A 比 Y 先选，A 不会嫉妒 Y。最后，A 也是不会嫉妒 X 的，因为即使 X 拥有了第二轮中的全部蛋糕，X 手里的蛋糕加起来也只是第一轮开始时 A 等分出来的其中一块蛋糕，这是不可能超过 A 的。这就说明了，三个人之间仍然不会有嫉妒发生，塞尔弗里奇–康威算法的确满足免嫉妒条件。

不过，塞尔弗里奇–康威算法只能在三人分蛋糕时使用，并不能扩展到人数更多的情况。对于人数更多的情况，免嫉妒分割问题更加困难，目前数学家们还没有找到一个比较可行的方案。正如数学家索尔·加芬克尔（Sol Garfunkel）所说，分蛋糕问题是 20 世纪数学研究中最重要的问题之一。直到现在，也还有一大群数学家正投身于分蛋糕问题之中，研究包括免嫉妒性在内的各种公平条件，致力于构造新的公平分割方案。



11. 中文自动分词算法

其实，我并不是数学专业的——大学时我一直在中文系的应用语言学专业读书。不过，我并不后悔当初的决定。正因为没在数学专业学习，我才能不以考试为目的地学习任何自己想学的数学知识，才能对数学有如此浓厚的兴趣。同时，应用语言学本身也是一门相当有意思的学问。在专业课上，我学到了很多计算机自动处理中文信息的算法，有的算法非常漂亮。自动分词可以说是信息处理的第一步，是这一领域中最简单最有趣的话题，在这里跟大家闲聊一下。

自动分词在互联网有着极其广泛的应用。当你在搜索引擎中搜索“软件使用技巧”时，搜索引擎通常会帮你找出同时含有“软件”、“使用”和“技巧”的网页，即使这三个词并没有连在一块儿。一个好的新闻网站通常会有“相关文章推荐”的功能，这也要依赖于自动分词的算法。不过，要想让计算机准确切分一句话，并不是那么容易。我就曾经看到过，某网站报道北京大学生怎么样怎么样，结果相关文章里列出的全是北京大学的新闻。这多半是分词算法错误地把标题中的“北京大学”当成了一个词。

那么，如何让计算机准确地切分一句话呢？

自动分词的主要困难在于分词歧义。“结婚的和尚未结婚的”，应该分成“结婚／的／和／尚未／结婚／的”，还是“结婚／的／和尚／未／结婚／的”？人来判断很容易，要交给计算机来处理就麻烦了。问题的关键就是，“和尚未”里的“和尚”也是一个词，“尚未”也是一个词，从计算机的角度看上去，两者似乎都有可能。对于计算机来说，这样的分词困境就叫做“交集型歧义”。

有时候，交集型歧义的“歧义链”有可能会更长。“中外科学名著”里，“中外”、“外科”、“科学”、“学名”、“名著”全是词，光从词库的角度来看，随便切几刀下去，得出的切分都是合理的。类似的例子数不胜数，“提高产品质量”、“鞭炮声响彻夜空”、



“努力学习语法规则”、“中国企业主要求解决”等句子都有这样的现象。在这些极端例子下，分词算法谁优谁劣可谓是一试便知。

最简单的，也是最容易想到的自动分词算法，便是“最大匹配法”了。也就是说，从句子左端开始，不断匹配最长的词（组不了词的单字则单独划开），直到把句子划分完。算法的理由很简单：人在阅读时也是从左往右逐字读入的，最大匹配法是与人的习惯相符的。而在大多数情况下，这种算法也的确能侥幸成功。不过，这种算法并不可靠，构造反例可以不费吹灰之力。例如，“北京大学生前来应聘”本应是“北京／大学生／前来／应聘”，却会被误分成“北京大学／生前／来／应聘”。

维护一个特殊规则表，可以修正一些很机械的问题，效果相当不错。例如，“不可能”要划分成“不／可能”，“会诊”后面接“断”、“疗”、“脉”、“治”时要把“会”单独切出，“的确切”后面是抽象名词时要把“的确切”分成“的／确切”，等等。

还有一个适用范围相当广的特殊规则，这个强大的规则能修正很多交集型歧义的划分错误。首先我们要维护一个一般不单独成词的字表，比如“民”、“尘”、“伟”、“习”等；这些字通常不会单独划出来，都要跟旁边的字一块儿组成一个词。在分词过程中，一旦发现这些字被孤立出来，都要重新考虑它与前面的字组词的可能性。例如，在用最大匹配法切分“为人民服务”时，算法会先划出“为人”一词，而后发现“民”字只能单独成词了。查表却发现，“民”并不能单独划出，于是考虑进行修正——把“为人”的“人”字分配给“民”字。碰巧这下“为”和“人民”正好都能成词，据此便可得出正确的划分“为／人民／服务”。

不过，上述算法归根结底，都是在像人一样从左到右地扫描文字。为了把问题变得更加形式化，充分利用计算机的优势，我们还有一种与人的阅读习惯完全不同的算法思路：把句子作为一个整体来考虑，从全局的角度评价一个句子划分方案的好坏。设计自动分词算法的问题，也就变成了如何评估分词方案优劣的问题。最初所用的办法就是，寻找词数最少的划分。注意，每次都匹配最长的词，得出的划分不见得是词数最少的，错误的贪心很可能会不慎错过一些更优的方案。因而，在有的情况下，最少词数法比最大匹配法效果更好。若用最大匹配法来划分，“独立自主和平等互利的原则”将被分成“独立自主／和平／等／互利／的／原则”，一共有 6 个词；但词数更少的方案则是“独立自主／和／平等互利／的／原则”，一共只有 5 个词。



当然，最少词数法也有出错的时候。“为人民办公益”的最大匹配划分和最少词数划分都是“为人／民办／公益”，而正确的划分则是“为／人民／办／公益”。同时，很多句子也有不止一个词数最少的分词方案，最少词数法并不能从中选出一个最佳答案。不过，把之前提到的“不成词字表”装备到最少词数法上，我们就有了一种简明而强大的算法：对于一种分词方案，里面有多少词，就罚多少分；每出现一个不成词的单字，就加罚一分。最好的分词方案，也就是罚分最少的方案。

这种算法的效果出人意料地好。“他说的确实在理”是一个很困难的测试用例，“的确”和“实在”碰巧也成词，这给自动分词带来了很大的障碍。但是“确”、“实”、“理”通常都不单独成词的，因此很多切分方案都会被扣掉不少分：

> 他／说／的／确实／在理（罚分：1+1+1+1+1=5）
> 他／说／的确／实／在理（罚分：1+1+1+2+1=6）
> 他／说／的确／实在／理（罚分：1+1+1+1+2=6）

正确答案胜出。

需要指出的是，这个算法并不需要穷举所有的划分可能。整个问题可以转化为图论中的最短路问题，利用一种叫做“动态规划”的技巧则会获得更高的效率。

算法还有进一步优化的余地。大家或许已经想到了，“字不成词”有一个程度的问题。“民”是一个不成词的语素，它是不能单独成词的。“鸭”一般不单独成词，但在儿歌童谣和科技语体中除外。“见”则是一个可以单独成词的语素，只是平时我们不常说罢了。换句话说，每个字成词都有一定的概率，每个词出现的概率也是不同的。

何不用每个词出现的概率，来衡量分词的优劣？于是我们有了一个更标准、更连续、更自动的改进算法，即最大概率法：先统计大量真实语料中各个词出现的概率，然后把每种分词方案中各词的出现概率乘起来作为这种方案的得分。最后，选出得分最高的方案，当作分词的结果。

以“有意见分歧”为例，让我们看看最大概率法是如何工作的。查表可知，在大量真实语料中，“有”、“有意”、“意见”、“见”、“分歧”的出现概率分别是 0.0181、0.0005、0.0010、0.0002、0.0001，因此“有／意见／分歧”的得分为1.8×10^{-9}，但“有意／见／分歧”的得分只有1.0×10^{-11}，正确方案完胜。

这里的假设是，用词造句无非是随机选词连在一块儿，是一个简单的一元过程。显然，这个假设理想得有点不合理，必然会有很多问题。考虑下面这句话：

这 / 事 / 的确 / 定 / 不 / 下来

但是概率算法却会把这个句子分成：

这 / 事 / 的 / 确定 / 不 / 下来

原因是，“的”字出现的概率太高了，它几乎总会从“的确”中挣脱出来。

其实，以上所有的分词算法都还有一个共同的大缺陷：它们虽然已经能很好地处理交集型歧义的问题，却完全无法解决另外一种被称为“组合型歧义”的问题。所谓组合型歧义，就是指同一个字串既可合又可分。比如说，“个人恩怨”中的“个人”就是一个词，“这个人”里的“个人”就必须拆开；“这扇门的把手”中的“把手”就是一个词，“把手抬起来”中的“把手”就必须拆开；“学生会宣传部”中的“学生会”就是一个词，“学生会主动完成作业”里的“学生会”就必须拆开。这样的例子非常多，“难过”、“马上”、“将来”、“才能”、“过人”、“研究所”、“原子能”都有此问题。究竟是合还是分，还得取决于它两侧的词语。到目前为止，所有算法对划分方案的评价标准都是基于每个词的固有性质的，完全不考虑相邻词语之间的影响，因而一旦涉及组合型歧义的问题，最大匹配、最少词数、概率最大等所有策略都不能实现具体情况具体分析。

于是，我们不得不跳出一元假设，把人类语言抽象成一个二元模型。对于任意两个词语 w_1 、w_2 ，统计在语料库中词语 w_1 后面恰好是 w_2 的概率 $P(w_1, w_2)$ 。这样便会生成一个很大的二维表。再定义一个句子的划分方案的得分为 $P(\phi, w_1) \cdot P(w_1, w_2) \cdot P(w_2, w_3) \cdot \cdots \cdot P(w_{n-1}, w_n)$ ，其中 $w_1, w_2, \cdots, w_n$ 依次表示分出的词，$P(\phi, w_1)$ 表示句子开头是 w_1 的概率。我们同样可以利用动态规划求出得分最高的分词方案。这真是一个天才的模型，这个模型一并解决了词类标注和语音识别等各类自然语言处理问题。

至此，中文自动分词算是有了一个漂亮而实用的算法。

但是，随便拿份报纸读读，你就会发现我们之前给出的测试用例都太理想了，简直就是用来喂给计算机的。在中文分词中，还有一个比分词歧义更令人头疼的东



西——未登录词。中文没有首字母大写，专名号也被取消了，这叫计算机如何辨认人名地名之类的东西？最近十年来，中文分词领域都在集中攻克这一难关。

在汉语的未登录词中，规律最强的要数中国人名了。根据统计，汉语姓氏大约有1000 多个，其中“王”、“陈”、“李”、“张”、“刘”五大姓氏的覆盖率高达 32%，前400 个姓氏覆盖率高达 99%。人名的用字也比较集中，“英”、“华”、“玉”、“秀”、“明”、“珍”六个字的覆盖率就有 10.35%，最常用的 400 字则有 90%的覆盖率。虽然这些字分布在包括文言虚词在内的各种词类里，但就用字的感情色彩来看，人名多用褒义字和中性字，少有不雅用字，因此规律性还是非常强的。根据这些信息，我们足以计算一个字符串能成为名字的概率，结合预先设置的阈值便能很好地识别出可能的人名。

可是，如何把人名从句子中切出来呢？换句话说，如果句子中几个连续字都是姓名常用字，人名究竟应该从哪儿取到哪儿呢？人名以姓氏为左边界，相对容易判定一些。人名的右边界则可以从下文的提示确定出来：人名后面通常会接“先生”、“同志”、“校长”、“主任”和“医生”等身份词，以及“是”、“说”、“报道”、“参加”、“访问”和“表示”等动作词。

但麻烦的情况也是有的。一些高频姓氏本身也是经常单独成词的常用字，例如“于”、“马”、“黄”、“常”和“高”等。很多反映时代性的名字也是本身就成词的，例如“建国”、“建设”、“国庆”和“跃进”等。更讨厌的就是那些整个名字本身就是常用词的人了，它们会彻底打乱之前的各种模型。如果分词程序也有智能的话，它一定会把所有叫“高峰”、“汪洋”和“黄莺”的人拖出去斩了。

还有那些恰好与上下文组合成词的人名，例如“费孝通向人大常委会提交书面报告”和“邓颖超生前使用过的物品”等，这就是最考验分词算法的时候了。

中国地名的用字就分散得多了。北京有一个地方叫“臭泥坑”，网上搜索“臭泥坑”，第一页全是“臭泥坑地图”和“臭泥坑附近酒店”之类的信息。某年《重庆晨报》刊登停电通知，上面赫然印着“停电范围包括沙坪坝区的犀牛屙屎和犀牛屙屎抽水”，读者纷纷去电投诉印刷错误。记者仔细一查，你猜怎么着，印刷并无错误，重庆真的就有叫“犀牛屙屎”和“犀牛屙屎抽水”的地方。

好在，中国地名数量有限，这是可以枚举的。中国地名委员会编写了《中华人民共和国地名录》，收录了从高原盆地到桥梁电站共 10 万多个地名，这让中国地名的识

别便利了很多。

外文人名和地名的用字则非常集中，识别起来也并不困难。

真正有些困难的就是识别机构名了，虽然机构名的后缀比较集中，但左边界的判断就有些难了。更难的就是品牌名了。如今各行各业大打创意战，品牌名可以说是无奇不有，而且经常本身就包含常用词，更是给自动分词添加了不少障碍。

最难识别的未登录词就是缩略语了。“高数”、“抵京”、“女单”、“发改委”、“北医三院”都是比较好认的缩略语，然而有些缩略语的含义连人都搞不清楚，又如何让计算机找出线索？你能猜到“人影办”是什么机构的简称吗？打死你都想不到，是“人工影响天气办公室”。

汉语中构造缩略语的规律很诡异，目前也没有一个定论。初次听到这个问题，几乎每个人都会做出这样的猜想：缩略语都是选用各个成分中最核心的字，比如“安全检查”缩成“安检”，“人民警察”缩成“民警”等等。不过，反例也是有的，“邮政编码”就被缩成了“邮编”，但“码”无疑是更能概括“编码”一词的。当然，这几个缩略语已经逐渐成词，可以加进词库了，但新近出现的或者临时构造的缩略语该怎么办，还真是个大问题。

说到新词，网络新词的大量出现才是分词系统真正的敌人。这些新词汇的来源千奇百怪，几乎没有固定的产生机制。要想实现对网络文章的自动分词，目前来看是相当困难的。革命尚未成功，分词算法还有很大的进步空间。

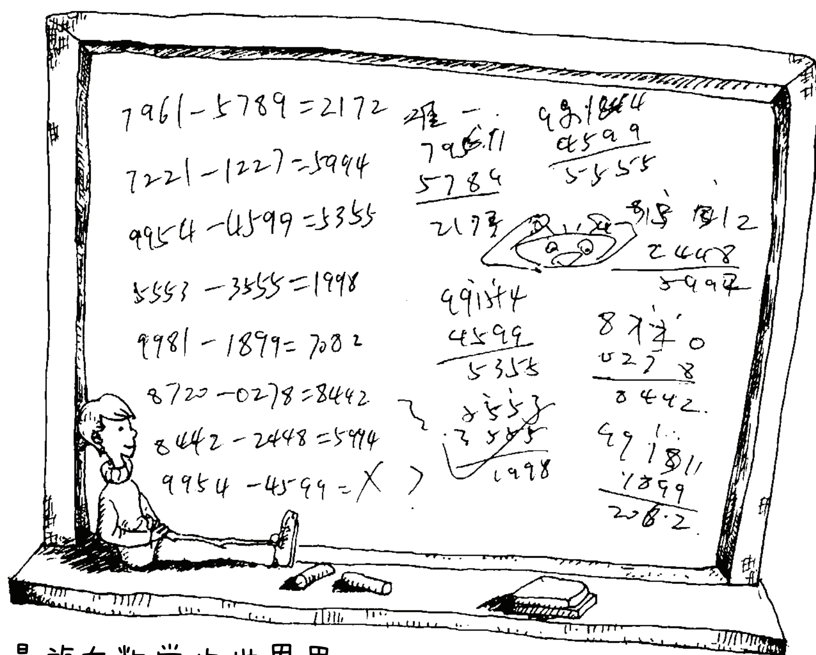


第二部分 数学之美

在数学发展的过程中，很多时候提出新的数学问题，开创新的数学领域，最初的动机并不是解释生活中的现象，而是因为它本身的美妙。数学世界里究竟有什么精彩之处，让数学家如此疯狂?



12. 让你立刻爱上数学的 8 个算术游戏



漫游在数学的世界里

文科背景的朋友们经常会问我一个问题：数学到底哪里有趣了，数学之美又在哪里？此时，我通常会讲一些简单而又深刻的算术游戏，让每个只会算术的人都能或多或少地体会到一些数学的美妙。如果你从小就被数学考试折磨，对数学一点好感都没有，那么我相信这一节内容会改变你的态度。

数字黑洞

任意选一个四位数（数字不能全相同），把所有数字从大到小排列，再把所有数



字从小到大排列，用前者减去后者得到一个新的数。重复对新得到的数进行上述操作，7 步以内必然会得到 6174。如果某一步计算的结果不足四位，那就在它前面添加 0，把它补成四位，再进行操作。例如，选择四位数 8080：

$$8800-0088=8712$$

$$8721-1278=7443$$

$$7443-3447=3996$$

$$9963-3699=6264$$

$$6642-2466=4176$$

$$7641-1467=6174$$

……

6174 这个“黑洞”就叫做卡布列克（Kaprekar）常数。对于三位数，也有一个数字黑洞，即 495。

特殊乘法的速算

如果两个两位数的十位数相同，个位数相加为 10，那么你可以立即说出这两个数的乘积。如果把这两个数分别写作 $\overline{AB}$ 和 $\overline{AC}$，那么它们的乘积的前两位就是 A 和 $A+1$ 的乘积，后两位就是 B 和 C 的乘积。

比如，47 和 43 的十位数相同，个位数之和为 10，因而它们乘积的前两位就是 $4\times(4+1)=20$，后两位就是 $7\times3=21$。也就是说，$47\times43=2021$。

类似地，$61\times69=4209$，$86\times84=7224$，$35\times35=1225$，等等。

这个速算方法背后的原因是，$(10x+y)(10x+(10-y))=100x(x+1)+y(10-y)$ 对任意 x 和 y 都成立。

翻倍，再翻倍！

将 123 456 789 翻倍，你会发现结果仍然是这 9 个数字的一个排列：

$$123\,456\,789\times2=246\,913\,578$$

我们再次将 246913578 翻倍，发现：

$$246\,913\,578\times2=493\,827\,156$$

结果依旧使用了每个数字各一次。这仅仅是一个巧合吗？我们继续翻倍：

493 827 156×2 = 987 654 312

神奇啊，一个很有特点的数 987 654 312，显然每个数字又只用了一次。

你或许会想，这下到头了吧，再翻倍就成 10 位数了。不过，请看：

987 654 312×2 = 1 975 308 624

又使用了每个数字各一次，只不过这一次加上了数字 0。再来？

1 975 308 624×2 = 3 950 617 248

恐怖了，又是每个数字各出现一次。

出现了这么多巧合之后我们开始怀疑，这并不是什么巧合，一定有什么简单的方法可以解释这种现象的。

但是，下面的事实让这个问题更加复杂了。到了第 6 次后，虽然仍然是 10 位数，但偏偏就在这时发生了意外：

3 950 617 248×2 = 7 901 234 496

看来，寻找一个合理的解释，并不是一件轻而易举的事情。

唯一的解

经典数字谜题：用 1 到 9 组成一个九位数，使得这个数的第一位能被 1 整除，前两位组成的两位数能被 2 整除，前三位组成的三位数能被 3 整除，以此类推，一直到整个九位数能被 9 整除。

没错，真的有这样猛的数：381 654 729。其中 3 能被 1 整除，38 能被 2 整除，381 能被 3 整除，一直到整个数能被 9 整除。这个数既可以用整除的性质一步步推出来，也可以利用计算机编程找到。

另一个有趣的事实是，在所有由 1 到 9 所组成的 362 880 个不同的九位数中，381 654 729 是唯一一个满足要求的数！

幻方之幻

一个“三阶幻方”是指把数字 1 到 9 填入 3×3 的方格，使得每一行、每一列以及两条对角线的 3 个数之和正好都相同。图 1 就是一个三阶幻方，每条直线上的 3 个数

之和都等于 15。

8	1	6
3	5	7
4	9	2

图 1

大家或许都听说过幻方这东西，但是并不知道幻方中的一些美妙的性质。例如，任意一个三阶幻方都满足，各行所组成的三位数的平方和，等于各行逆序所组成的三位数的平方和。对于上图中的三阶幻方，就有

$$816^2+357^2+492^2=618^2+753^2+294^2$$

利用线性代数，我们可以证明这个结论。

天然形成的幻方

从 $\frac{1}{19}$ 到 $\frac{18}{19}$ 这 18 个分数的小数循环节长度都是 18。像图 2 那样把这 18 个循环节排成一个 18×18 的数字阵，这将恰好构成一个幻方——每一行、每一列和两条对角线上的数字之和都是 81。[①]

1/19 = 0.	0	5	2	6	3	1	5	7	8	9	4	7	3	6	8	4	2	1	...
2/19 = 0.	1	0	5	2	6	3	1	5	7	8	9	4	7	3	6	8	4	2	...
3/19 = 0.	1	5	7	8	9	4	7	3	6	8	4	2	1	0	5	2	6	3	...
4/19 = 0.	2	1	0	5	2	6	3	1	5	7	8	9	4	7	3	6	8	4	...
5/19 = 0.	2	6	3	1	5	7	8	9	4	7	3	6	8	4	2	1	0	5	...
6/19 = 0.	3	1	5	7	8	9	4	7	3	6	8	4	2	1	0	5	2	6	...
7/19 = 0.	3	6	8	4	2	1	0	5	2	6	3	1	5	7	8	9	4	7	...
8/19 = 0.	4	2	1	0	5	2	6	3	1	5	7	8	9	4	7	3	6	8	...
9/19 = 0.	4	7	3	6	8	4	2	1	0	5	2	6	3	1	5	7	8	9	...
10/19 = 0.	5	2	6	3	1	5	7	8	9	4	7	3	6	8	4	2	1	0	...
11/19 = 0.	5	7	8	9	4	7	3	6	8	4	2	1	0	5	2	6	3	1	...
12/19 = 0.	6	3	1	5	7	8	9	4	7	3	6	8	4	2	1	0	5	2	...
13/19 = 0.	6	8	4	2	1	0	5	2	6	3	1	5	7	8	9	4	7	3	...
14/19 = 0.	7	3	6	8	4	2	1	0	5	2	6	3	1	5	7	8	9	4	...
15/19 = 0.	7	8	9	4	7	3	6	8	4	2	1	0	5	2	6	3	1	5	...
16/19 = 0.	8	4	2	1	0	5	2	6	3	1	5	7	8	9	4	7	3	6	...
17/19 = 0.	8	9	4	7	3	6	8	4	2	1	0	5	2	6	3	1	5	7	...
18/19 = 0.	9	4	7	3	6	8	4	2	1	0	5	2	6	3	1	5	7	8	...

图 2

① 严格意义上说它不算幻方，因为方阵中有相同的数字。

一个小魔术

在一张纸上并排画 11 个小方格，叫你的好朋友背对着你（让你看不到他在纸上写什么），在前两个方格中随便填两个 1 到 10 之间的数。从第 3 个方格开始，在每个方格里填入前两个方格里的数之和。让你的朋友一直算出第 10 个方格里的数。假如你的朋友一开始填入方格的数是 7 和 3，那么前 10 个方格里的数分别是：

7	3	10	13	23	36	59	95	154	249

现在，叫你的朋友报出第 10 个方格里的数，稍作计算你便能猜出第 11 个方格里的数应该是多少。你的朋友会非常惊奇地发现，把第 11 个方格里的数计算出来，所得的结果与你的预测一模一样！

其实，仅凭借第 10 个数来推测第 11 个数的方法非常简单，你需要做的仅仅是把第 10 个数乘以 1.618，得到的乘积就是第 11 个数了。在上面的例子中，由于 $249\times1.618=402.882\approx403$，因此你可以胸有成竹地断定，第 11 个数就是 403。而事实上，154 与 249 相加真的就等于 403。

其实，不管最初两个数是什么，按照这种方式加下去，相邻两数之比总会越来越趋近于 1.618——这个数正是传说中的“黄金分割”。

3 个神奇的分数

$\frac{1}{49}$ 化成小数后等于 0.0204081632…，把小数点后的数字两位两位断开，前五个数依次是 2、4、8、16、32，每个数正好都是前一个数的两倍。

$\frac{100}{9899}$ 等于 0.01010203050813213455…，两位两位断开后，得到的正好是著名的斐波那契（Fibonacci）数列 1, 1, 2, 3, 5, 8, 13, 21, …，数列中的每一个项都是它前面两个项之和。

而 $\frac{100}{9801}$ 则等于 0.0102030405060708091011121314151617181920212223…。

利用组合数学中的“生成函数”可以完美地解释这些现象产生的原因。



13. 最折磨人的数学未解之谜

数学之美不但体现在漂亮的结论和精妙的证明上，那些尚未解决的数学问题也有让人神魂颠倒的魅力。和哥德巴赫猜想、黎曼假设不同，有些悬而未解的问题趣味性很强，"数学性"却非常弱，乍看上去并没有触及深刻的数学理论，似乎是一道可以被瞬间秒杀的数学趣题，让数学爱好者们"不找到一个巧解就不爽"；但令人称奇的是，它们的困难程度却不亚于那些著名的数学猜想，这或许比各个领域中艰深的数学难题更折磨人吧。

$3x+1$问题

从任意一个正整数开始，重复对其进行下面的操作：如果这个数是偶数，把它除以 2；如果这个数是奇数，则把它扩大到原来的 3 倍后再加 1。序列是否最终总会变成 4, 2, 1, 4, 2, 1, … 这种循环？

这个问题可以说是一个"坑"——乍看之下，问题非常简单，突破口很多，于是数学家们纷纷往里面跳；殊不知进去容易出来难，不少数学家到死都没把这个问题搞出来。已经中招的数学家不计其数，这可以从 $3x+1$ 问题的各种别名看出来：$3x+1$ 问题又叫科拉兹（Collatz）猜想、叙拉古（Syracuse）问题、角谷猜想、哈斯（Hasse）算法和乌拉姆（Ulam）问题等。后来，由于命名争议太大，干脆让谁都不沾光，直接叫做 $3x+1$ 问题算了。

$3x+1$ 问题不是一般地困难。这里举一个例子说明数列收敛有多么没规律。从 26 开始算起，10 步就掉入了"421 陷阱"：

26, 13, 40, 20, 10, 5, 16, 8, 4, 2, 1, 4, 2, 1, …

但是，从 27 开始算起，数字会一路飙升到几千之大，你很可能会一度认为它脱离了“421 陷阱”。但是，经过上百步运算后，它还是跌了回来：

27, 82, 41, 124, 62, 31, 94, 47, 142, 71, 214, 107, 322, 161, 484, 242, 121, 364, 182, 91, 274, 137, 412, 206, 103, 310, 155, 466, 233, 700, 350, 175, 526, 263, 790, 395, 1186, 593, 1780, 890, 445, 1336, 668, 334, 167, 502, 251, 754, 377, 1132, 566, 283, 850, 425, 1276, 638, 319, 958, 479, 1438, 719, 2158, 1079, 3238, 1619, 4858, 2429, 7288, 3644, 1822, 911, 2734, 1367, 4102, 2051, 6154, 3077, 9232, 4616, 2308, 1154, 577, 1732, 866, 433, 1300, 650, 325, 976, 488, 244, 122, 61, 184, 92, 46, 23, 70, 35, 106, 53, 160, 80, 40, 20, 10, 5, 16, 8, 4, 2, 1, 4, 2, 1, …

196 问题

如果一个数正读反读都一样，我们就把它叫做“回文数”。随便选一个数，不断加上把它反过来写之后得到的数，直到得出一个回文数为止。例如，所选的数是 67，两步就可以得到一个回文数 484：

67 + 76 = 143

143 + 341 = 484

把 69 变成一个回文数则需要四步：

69 + 96 = 165

165 + 561 = 726

726 + 627 = 1353

1353 + 3531 = 4884

89 的“回文数之路”则特别长，要到第 24 步才会得到第一个回文数，8 813 200 023 188。

大家或许会想，不断地“一正一反相加”，最后总能得到一个回文数，这当然不足为奇了。事实似乎也确实是这样的——对于几乎所有的数，按照规则不断加下去，迟早会出现回文数。不过，196 却是一个相当引人注目的例外。数学家们已经用计算机算到了 3 亿多位数，都没有产生过一次回文数。从 196 出发，究竟能否加出回文数来？196 究竟特殊在哪儿？这至今仍是个谜。



随机 01 串的最长公共子序列

如果从数字序列 A 中删除一些数字就能得到数字序列 B，我们就说 B 是 A 的子序列。例如，110 是 010010 的子序列，但不是 001011 的子序列。两个序列的“公共子序列”有很多，其中最长的那个就叫做“最长公共子序列”。

随机产生两个长度为 n 的 01 序列，其中数字 1 出现的概率是 p，数字 0 出现的概率是 $1-p$。用 $C_p(n)$ 来表示它们的最长公共子序列的长度，用 C_p 来表示当 n 无穷大时 $\frac{C_p(n)}{n}$ 的极限值。

关于 C_p 的存在性，有一个非常巧妙的证明；然而，这个证明仅仅说明了 C_p 存在，它没有给计算 C_p 带来任何有用的提示。

即使是 $C_{1/2}$ 的值，也没人能成功算出来。迈克尔·斯蒂尔（Michael Steele）猜想 $C_{1/2}=\frac{2}{1+\sqrt{2}}\approx 0.828\,427$。后来，瓦克拉夫·克沃特尔（Vaclav Chvátal）和戴维·桑科夫（David Sankoff）证明了 $0.773\ 911<C_{1/2}<0.837\ 623$，看上去迈克尔·斯蒂尔的猜想似乎很可能是对的。2003 年，乔治·利克（George Lueker）证明了 $0.7880<C_{1/2}<0.8263$，推翻了迈克尔·斯蒂尔的猜想。

更糟的是，“当 $p=\frac{1}{2}$ 时 C_p 达到最小”似乎是一件很靠谱的事，但这个结论也无人能证明。

克拉科斯基数列

克拉科斯基（Kolakoski）数列是一个仅由 1 和 2 构成的数列，其中头 100 个数是：

1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 1, 1, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 1,
2, 2, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 1, 1, 2, 1, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1,
2, 1, 1, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 2, 1, 1, 2, 2, ⋯

如果我们把连续的相同数看做一组的话，整个数列的定义就只有两句话：$a(1)=1$，$a(n)$ 表示第 n 组数的长度。注意，有了这几个条件，整个序列就已经唯一地确定了！$a(1)=1$ 就表明第一组数只有一个数（也就是它自己），因此下一个数必须要换成 2，也就是 $a(2)=2$；而 $a(2)=2$ 又说明第二组数（也就是 $a(2)$ 所在的这组数）有两个数，

因此 $a(3)$ 也等于 2；而 $a(3)=2$ 就表明第三组数的长度为 2，即数列接下来要有两个 1，等等。也就是说，这个数列完全是“自生成”的。更酷的说法则是，如果把每一组数用它的长度来替换，得到的仍然是这个数列本身。

关于克拉科斯基数列，我们知道些什么？很少。贝诺瓦·克罗伊特（Benoit Cloitre）发现，这个数列可以用递归式 $a(a(1)+a(2)+\cdots+a(k))=\frac{3+(-1)^k}{2}$ 来表达。德金（F. M. Dekking）证明了一个看上去更妙的结论：去掉数列最前面的 1，剩下的部分可以从 22 开始，每次按 22→2211，21→221，12→211，11→21 的规则两位两位地对数列进行替换，并不断迭代产生。不过，这些发现都不足以让我们更加深入地了解克拉科斯基数列。

克拉科斯基数列的第 n 项有非递归的公式吗？目前我们还不知道。已经出现过的数字串今后都还会再次出现吗？目前我们也不知道。还有，我们有理由猜想，数列中 1 和 2 的个数各占一半。图 1 显示的就是数列前 n 项中数字 1 所占的比例，可见我们的猜想很可能是对的。

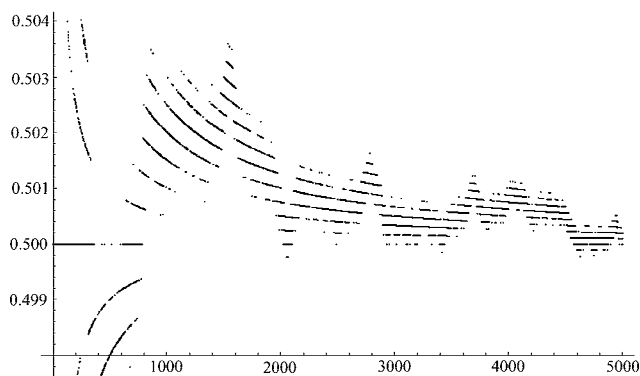


图　1

不过，目前还没有人能够证明这一点。而最近的一些研究表明，数字 1 的比例很可能不是 $\frac{1}{2}$。当然，还有第三种可能——这个极限可能根本不存在。

吉尔布雷思猜想

从小到大依次列出所有的质数：



2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, …

求出相邻两项之差：

1, 2, 2, 4, 2, 4, 2, 4, 6, 2, …

现在，再次求出所得序列中相邻两项之差，又会得到一个新的序列：

1, 0, 2, 2, 2, 2, 2, 2, 4, …

重复对所得序列进行这样的操作，我们还可以依次得到

1, 2, 0, 0, 0, 0, 0, 2, …
1, 2, 0, 0, 0, 0, 2, …
1, 2, 0, 0, 0, 2, …
1, 2, 0, 0, 2, …

大家会发现一个有趣的规律：每行序列的第一个数都是 1。

某日，数学家诺曼·吉尔布雷思（Norman L. Gilbreath）闲得无聊，在餐巾上不断对质数序列求差，于是发现了上面这个规律。吉尔布雷思的两个学生对前 64 419 行序列进行了检验，发现这个规律始终成立。1958 年，吉尔布雷思在一个数学交流会上提出了他的发现，吉尔布雷思猜想由此诞生。

这个规律如此之强，很少有人认为猜想不成立。1993 年，安德鲁·奥德里兹科（Andrew Odlyzko）对 10 000 000 000 000 以内的质数（也就是 346 065 536 839 行）进行了检验，也没有发现反例。

不过，这一看似简单的问题，几十年来硬是没人解决。

辛马斯特猜想

图 2 所示为杨辉三角[①]，其中数字 1 出现了无穷多次。除了数字 1 以外，哪个数

① 又叫做帕斯卡（Pascal）三角，是一个由正整数构成的三角形数阵，其生成规律非常简单：每行左右两头的数都是 1，中间的数都是它左上角的数和右上角的数之和。在代数和组合数学中，杨辉三角都有着非常重要的意义。如果把首行称做“第 0 行”（因而第二行就叫做“第 1 行”），把每行的头一个数叫做“第 0 个数”（因而第二个数才是“第 1 个数”）的话，那么杨辉三角第 m 行的第 n 个数就等于 $(1+x)^m$ 的展开式中的 n 次项系数，也是从 m 个不同物体中取出其中 n 个物体的方案数 C_m^n 。在后文中，我们还会提到杨辉三角。

字出现的次数最多呢？6 出现了 3 次，不过不算多。10 出现了 4 次，不过也不算多。120 出现了 6 次，算多了吧？还不算多。目前已知的出现次数最多的数是 3003，它同时等于 C_{3003}^{1}、C_{3003}^{3002}、C_{78}^{2}、C_{78}^{76}、C_{15}^{5}、C_{15}^{10}、C_{14}^{6}、C_{14}^{8}，在杨辉三角中出现了 8 次。有没有出现次数更多的数，目前仍然是一个未解之谜。

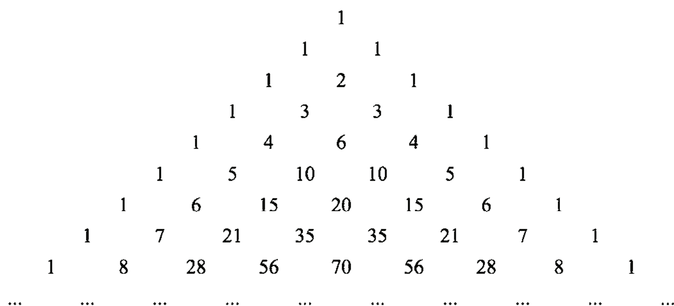


图 2

真正精彩的来了。如果把正整数 $a>1$ 在杨辉三角中出现的次数记做 $N(a)$，那么函数 $N(a)$ 是以什么级别上涨的呢？1971 年，戴维·辛马斯特（David Singmaster）证明了 $N(a)=O(\log a)$，即 $N(a)$ 最多是以对数级别上涨的。他同时猜想 $N(a)=O(1)$，即 $N(a)$ 有一个上限。这也就是辛马斯特猜想。由于我们一直没能找到出现次数超过 8 的数，因而这个上界很可能就是 8。不过，辛马斯特猜测这个上界更可能是 10 或者 12。

保罗·埃尔德什（Paul Erdős）[①]认为，辛马斯特的猜想很可能是正确的，但证明起来会非常困难。目前最好的结果是，$N(a)=O\left(\dfrac{\log a\cdot\log\log\log a}{(\log\log a)^{3}}\right)$。

孤独的赛跑者

有一个环形跑道，总长为 1 个单位。n 个人从跑道上的同一位置出发，沿着跑道顺时针一直跑下去。每个人的速度都是固定的，但不同人的速度不同。证明或推翻，对于每一个人，总会有一个时刻，他与其他所有人的距离都不小于 $\dfrac{1}{n}$。

① 匈牙利数学家，在数学界极其活跃，一生中与数百人合作，发表过 1525 篇数学论文，是目前发表论文数最多的数学家。埃尔德什研究过很多数学谜题，并给出了异常漂亮的解答；但同时，他也遇到了很多至今仍未解决的数学难题，并立下了大大小小的悬赏。不过他坚信，上帝手中有一本书，书中记载了所有数学定理最精妙的证明。记住埃尔德什这个名字，我们后面还会反复提到他。



这个问题是由威尔斯（J. M. Wills）在1967年提出的。乍看上去，这个问题无异于其他各种非常巧妙的初等组合数学问题，但不可思议的是，这个问题竟然直到现在仍没彻底解决。

当$n=2$时，由于两人的速度不同，因此到了某个时刻，他们必然会位于环形跑道的两个对称位置上，他们到对方的距离都恰好等于$\frac{1}{2}$，可见$n=2$时命题是成立的。此后，数学家们先后证明了$n=3$、$n=4$、$n=5$和$n=6$的情形。直觉上，对于更大的n，结论也应该成立，不过尚未有人证明。

双倍困难的排序问题

有n个盒子，从左至右依次编号为$1, 2, \cdots, n$。第1个盒子里放两个编号为n的小球，第2个盒子里放两个编号为$n-1$的小球，以此类推，第n个盒子里放两个编号为1的小球。每次你可以在相邻两个盒子中各取一个小球，交换它们的位置。为了把所有小球放进正确的盒子里，最少需要几次交换？

为了说明这个问题背后的陷阱，我们不妨先拿$n=5$的情况做个例子。首先，如果每个盒子里只有一个球，问题就变成了经典的排序问题了：只能交换相邻元素，如何最快地把5, 4, 3, 2, 1变成1, 2, 3, 4, 5？如果一个数列中前面的某个数反而比后面的某个数大，我们就说这两个数是一个“逆序对”。显然，初始情况下所有数对都是逆序对，$n=5$时逆序对共有10个。我们的目的就是要把这个数目减少到0。而交换两个相邻的数只能消除一个逆序对，因此10次交换是必需的。

不过，题目中每个盒子里有两个球，那么是不是必须要交换20次才行呢？错！下面这种做法可以奇迹般地在15步之内完成排序。

55, 44, 33, 22, 11

54, 54, 33, 22, 11

54, 43, 53, 22, 11

54, 43, 32, 52, 11

54, 43, 32, 21, 51

54, 43, 21, 32, 51



54, 31, 42, 32, 51

41, 53, 42, 32, 51

41, 32, 54, 32, 51

41, 32, 42, 53, 51

41, 32, 42, 31, 55

41, 32, 21, 43, 55

41, 21, 32, 43, 55

11, 42, 32, 43, 55

11, 22, 43, 43, 55

11, 22, 33, 44, 55

第一次看上去似乎很不可思议，但细想一下还是能想明白的：同一个盒子里能够放两个数，确实多了很多新的可能。如果左边盒子里的某个数比右边某个盒子里的数大，我们就说这两个数构成一个逆序对；但如果两个不同的数在同一个盒子里，我们就把它们视作半个逆序对。现在让我们来看看，一次交换最多能消除多少个逆序对。假设某一步交换把 ab 和 cd 变成了 ac 和 bd，最好的情况就是 bc 这个逆序对彻底消除了，同时 ac 和 bd 两个逆序对消除了一半，ab 和 cd 两个（已经消除了一半的）逆序对也消除了一半，因此一次交换最多可以消除 3 个逆序对。由于一开始每个盒子里的两个相同的数都会在中间的某个时刻分开来，最后又会合并在一起，因此我们可以把初始时两个相同的数也当做一个逆序对。这样的话，初始时每两个数都是逆序对，n 个盒子里将产生 C_{2n}^2 个逆序对。自然，我们至少需要 $C_{2n}^2/3$ 步才能完成排序。当 $n=5$ 时，$C_{2n}^2/3=15$，这就说明了上面给出的 $n=5$ 的排序方案是最优的。

这个分析太巧妙了，实在是让人拍案叫绝。只可惜，这个下界并不是总能达到的。当 $n=6$ 时，上述分析得出的下界是 22 步，但计算机穷举发现没有 23 步交换是不行的。于是，这个问题又变成了一个诱人的坑，至今仍未被填上。

曲线的内接正方形

证明或推翻，在平面中的任意一条简单封闭曲线上，总能找到 4 个点，它们恰能组成一个正方形。



这样一个看上去如此基本的问题，竟然没有被解决！目前，对于充分光滑的曲线，似乎已经有了肯定的结论；但对于任意曲线来说，这仍然是一个悬而未解的问题。平面上的曲线无奇不有，说不准我们真能精心构造出一种不满足要求的怪异曲线。

多面体的展开

证明或推翻，总可以把一个凸多面体沿着棱剪开，展开成一个简单的（也就是不与自身相交的）平面多边形。

这是一个看上去很“自然”的问题，或许大家在玩弄各种纸制包装盒的时候，就已经思考过这个问题了。现在，人们已经找到了不满足条件的凹多面体，也就是说存在凹多面体，无论怎样展开它都会不可避免地得到与自身重叠的平面多边形。同时，确实也存在一些凸多面体，按照某种方式展开它后，会得到与自身重叠的平面多边形。不过，对于某个凸多面体，任何一种方法都不能把它展开到一个平面上，这听上去似乎不大可能；然而，在数学上这一点却一直没被证明。

线段距离的频数

n 个点一共可以确定 C_n^2 条线段，而这个数正好等于 $1+2+3+\cdots+(n-1)$ ——在本书的第四部分，大家将会看到 $C_n^2=1+2+3+\cdots+(n-1)$ 的一个非常漂亮的证明。于是我们想问，是否对于任意正整数 n，总能找出平面上处于一般位置（任意三点不共线、任意四点不共圆）的 n 个点，使得其中有一种长度的线段恰好出现了 1 次，有一种长度的线段恰好出现了 2 次，等等，一直到有一种长度的线段恰好出现了 $n-1$ 次？

当 $n=3$ 时，任意一个不是等边三角形的等腰三角形都满足要求。当 $n=4$ 时，可以先把其中三个点摆成一个等边三角形，第四个点则放在某一边的中垂线上，但不要让它与等边三角形的中心重合，于是就得到了图 3 所示的图形。这个图中线段的长度有 3 种，它们各出现了 1 次、2 次、3 次，因而正好满足要求。

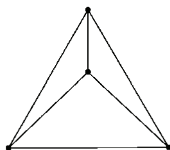


图　3

当 $n=5$ 时，这样的图形还存在吗？受很多与维度有关的几何命题的影响，或许很多人会认为，这样的图形只在更高维的空间中才存在吧。其实不然，在平面中也存在 $n=5$ 的解。图 4 就是一个简单的构造：$\triangle ABC$ 为等边三角形，O 为其中心，再以 A 为圆心，AB 为半径作弧，OB 的中垂线与这段弧相交于点 D。容易看出，$AB=BC=AC=AD$，$AO=BO=CO$，$DB=DO$，只有 CD 的长度是独一无二的。这就是一个满足要求的图形。

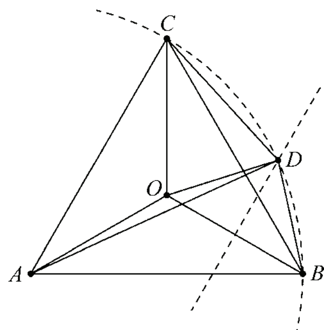


图　4

不但如此，人们还找到了 n 等于 6、7、8 时的解。图 5 就是 $n=8$ 时的一个解，大家可以验证一下。不过，继续往前探索的路偏偏就卡在了这里。对于 $n>8$ 的情况究竟是否有解，目前还没有一个定论。

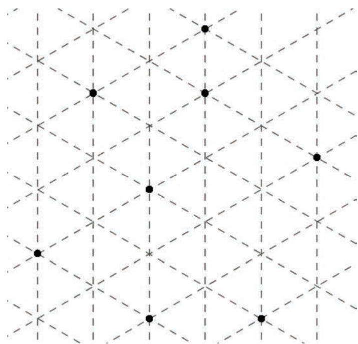


图　5

数学家们似乎更倾向于相信，当 n 足够大时，总会发生无解的情况。埃尔德什悬赏 50 美元征求当 n 足够大时问题无解的证明，同时悬赏 500 美元征求对任意 n 都适用



的构造解。

库斯纳猜想

很多城市的交通系统都是由大量横纵街道交错构成的，纽约的曼哈顿区就是最为典型的例子。因此，在估算两地之间的距离时，我们往往不会直接去测量两地之间的直线距离，而会去考虑它们在横纵方向上一共相距多少个街区。在数学中，我们就把平面上两个点的横坐标之差与纵坐标之差的和叫做这两点之间的曼哈顿距离。例如，$(0,0)$ 和 $(3,4)$ 两点间的直线距离是 5，但曼哈顿距离则是 7。

这个定义可以很自然地推广到 n 维空间中去。定义 n 维空间中 $P(p_1,p_2,\cdots,p_n)$ 和 $Q(q_1,q_2,\cdots,q_n)$ 两点之间的曼哈顿距离为 $|p_1-q_1|+|p_2-q_2|+\cdots+|p_n-q_n|$，直观地说，就是在 n 维网格中从 P 到 Q 的最短路径长度。某日，网友木遥[①]告诉了我一个与此相关的数学未解之谜：在 n 维空间中，最多可以有多少个曼哈顿距离两两相等的点？

容易看出，这样的点至少可以有 $2n$ 个，例如三维空间中 $(1,0,0)$、$(-1,0,0)$、$(0,1,0)$、$(0,-1,0)$、$(0,0,1)$、$(0,0,-1)$ 就是满足要求的 6 个点。大家肯定会想，这应该就是点数最多的方案了吧？不过，真要证明起来可没那么容易。1983 年，罗伯特·库斯纳（Robert Kusner）猜想，n 维空间中曼哈顿距离两两相等的点最多也只能有 $2n$ 个，这也就是现在所说的库斯纳猜想。目前人们已经证明，当 $n \leqslant 4$ 时，库斯纳猜想是正确的。当 $n>4$ 时呢？虽然大家相信这个猜想也应该是正确的，但还没有人能够证明。

有趣的是，在很多其他的度量空间下，同类型的问题却并没有这么棘手。如果把距离定义为标准的直线距离，那么 n 维空间中显然最多有 $n+1$ 个等距点；如果把距离定义为切比雪夫（Chebyshev）距离（即所有 $|p_i-q_i|$ 中的最大值），问题的解则是 2^n，即 n 维坐标系中单位立方体的 2^n 个顶点。一旦换作曼哈顿距离，问题就迟迟不能解决，这还真有些出人意料。

Thrackle 猜想

在纸上画一些点，再画一些点与点之间的连线，我们就把所得的图形叫做一个“图”。如果一个图的每根线条都与其他所有线条恰好相交一次（顶点处相接也算相

① 木遥的博客地址为：http://blog.farmostwood.net。



交），那么就把这个图叫做一个 thrackle。图 6 显示的就是三个满足要求的 thrackle，注意到它们的线条数量都没有超过顶点的数量。问，是否存在线条数大于顶点数的 thrackle？

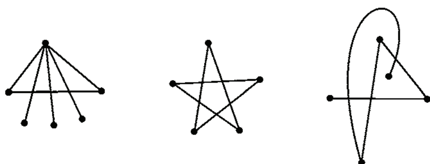


图　6

这个问题是由数学家约翰·康威提出来的。这明显又是一个坑，看到这个问题谁都想试试，然后就纷纷崩溃掉。康威悬赏 1000 美元征解，可见这个问题有多么不容易。目前已知的最好的结果是，一个 thrackle 的线条数不会超过顶点数的 $\frac{167}{117}$ 倍。

拉姆齐问题

有这么一个定理：6 个人参加一个聚会，其中某些人之间握过手，那么一定存在 3 个人互相之间都握过手，或者 3 个人互相之间都没握过手。我们可以借助鸽笼原理[①]很快证明这个结论。选出其中一个人 A，然后把剩下的 5 个人分成 2 组，和 A 握过手的，以及没和 A 握过手的。显然，其中一组至少有 3 个人。不妨假设和 A 握过手的那一组至少有 3 个人吧。（在另一种情况下，下述推理同样适用。）把这一组里的 3 个人分别记作 B、C、D（如果这一组的人数大于 3，任意选 3 个人就行了）。如果 B、C、D 这 3 个人之间有 2 个人握过手，那么这 2 个人和 A 就成了互相之间握过手的 3 人组；如果 B、C、D 这 3 个人之间都没握过手，那么他们本身就成了互相之间都没握过手的 3 人组。

1930 年，英国数学家弗兰克·拉姆齐（Frank Ramsey）证明了一个更强的结论：给定两个正整数 r 和 s，总能找到一个 n，使得一场 n 人聚会中，或者存在 r 个人互

① 假设有 n 只鸽子飞回 m 个笼子，如果 $n > m$ 的话，那么一定有至少一个笼子，它里面有不止一只鸽子。事实上，至少有一个笼子，它里面有不少于 $\left\lceil \frac{n}{m} \right\rceil$ 只鸽子，其中 $\lceil x \rceil$ 表示大于等于 x 的最小整数。鸽笼原理是组合数学中的一个重要工具，今后我们还会用到。



相之间都握过手，或者存在 s 个人互相之间都没握过手。我们把满足条件的最小的 n 记作 $R(r,s)$ 。

前面我们已经证明了，6 个人足以产生互相都握过手的 3 个人或者互相都没握过手的 3 个人，也就是说 $R(3,3) \leqslant 6$ 。但 5 个人是不够的，比方说只有 A 和 B、B 和 C、C 和 D、D 和 E、E 和 A 之间握手，容易看出不管选哪 3 个人，握过手的和没握过手的总是并存的 。因此， $R(3,3)$ 精确地等于 6。

求出 $R(r,s)$ 的精确值出人意料地难。目前已经知道 $R(4,4)=18$ ，但对于 $R(5,5)$ ，我们只知道它介于 43 到 49 之间，具体的值至今仍未求出来。如果要用计算机硬求 $R(5,5)$ ，则计算机需要考虑的情况数大约在 10^{300} 这个数量级，这是一个不可能完成的任务。而 $R(6,6)$ 就更大了，目前已知它在 102 到 165 的范围内。它的准确值是多少，恐怕我们永远都不可能知道了。

埃尔德什曾经说过，假如有一支异常强大的外星人军队来到地球，要求人类给出 $R(5,5)$ 的准确值，否则就会摧毁地球，那么他建议，此时我们应该集结全世界所有数学家的智慧和全世界所有计算机的力量，试着求出 $R(5,5)$ 来。但是，假如外星人要求人类给出 $R(6,6)$ 的准确值，那么他建议，我们应该试着摧毁外星人军队。

维恩图并不简单

给定 n 个集合后，每一个元素都拥有了自己的位置。比方说，若有“质数”、“两位数”、“个位是 3 的数” 这 3 个集合，则 31 就只属于前两个集合，而 102 则不属于任一个集合。我们往往会像图 7 左边那样，把这些集合抽象成一个个圆圈并画在同一平面上，然后把各个元素填入图中适当的区域，从而直观地展示出每个元素的所属情况。这样的图就叫做维恩（Venn）图。为了展示出由这 n 个集合产生的所有关系，维恩图需要有 2^n 个区域（包括最外面的那个区域）。

画惯了 3 个集合的维恩图，很多人都会认为，像图 7 右边那样把 4 个圆圈画成一朵花，就是 4 个集合的维恩图了。其实这是不对的——4 个圆只能产生 14 个区域，而 4 个集合将会交出 16 种情况。如果把 4 个圆圈像图 7 右边那幅图一样排列，就少了 2 个区域：只属于左下角的圆和右上角的圆的区域，以及只属于左上角的圆和右下角的圆的区域。

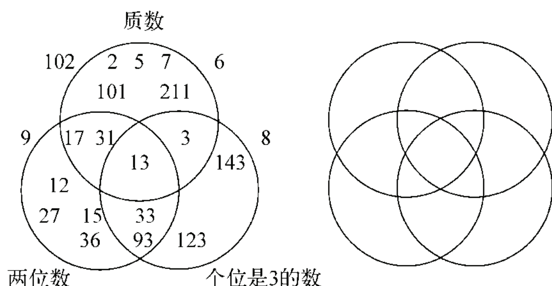


图　7

那么，是不是 4 个集合的维恩图就没法画了呢？也不是。如果你不是一个完美主义者，你可以像图 8 那样，把 3 个集合的维恩图扩展到 4 个集合；虽然看上去非常不美观，但是从拓扑学的角度来说，只要逻辑上正确无误，谁管它画得圆不圆呢。

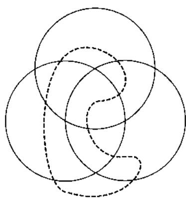


图　8

大家会自然而然地想到一个问题：右边这个图是否还能继续扩展成 5 个集合的维恩图呢？更一般地，是否随便什么样的 n 个集合的维恩图都可以扩展到 $n+1$ 个集合呢？

令人难以置信的是，这个问题竟然还没被解决！事实上，对满足各种条件的维恩图的研究是一个经久不衰的话题，与维恩图相关的猜想绝不止这一个。

遍历所有的“中间子集”

证明或推翻，你可以通过每次添加或者删除一个元素，循环遍历集合 $\{1,2,\cdots,2n+1\}$ 的所有大小为 n 或 $n+1$ 的子集。例如，当 $n=2$ 时，你可以通过以下路径循环遍历 $\{1,2,3,4,5\}$ 的所有包含 2 个元素或者 3 个元素的子集：

$\{1, 2\} \to \{1, 2, 3\} \to \{1, 3\} \to \{1, 3, 4\} \to \{1, 4\} \to \{1, 2, 4\} \to \{2, 4\} \to \{2, 4, 5\} \to \{4, 5\} \to \{1, 4, 5\} \to \{1, 5\} \to \{1, 3, 5\} \to \{3, 5\} \to \{3, 4, 5\} \to \{3, 4\} \to \{2, 3, 4\} \to \{2, 3\} \to \{2, 3, 5\} \to \{2, 5\} \to \{1, 2, 5\} \to \{1, 2\}$



看完上面的这段内容，我可以想象你已经有一种克制不住的冲动，拿起铅笔和草稿纸，或者跑到电脑前，开始寻找 n 不大时的规律。这可以说是本文的所有问题中最大的一个坑了——这个问题极具诱惑性，任何人第一次看到这个问题时都会认为存在一种对所有 n 都适用的构造解，于是众人一个接一个地往坑里跳，拦都拦不住。

几乎没有人认为这个猜想是错误的。目前计算机已经验证了，当 $n \leqslant 17$ 时，猜想都是成立的。从已有数据来看，随着 n 的增加，遍历这些子集的方案数不但也随之增加，而且增长得非常快。到了某个 n，方案数突然跌到了 0，这明显是一件极不可能发生的事。但是，几十年过去了，却没有人能够证明它！

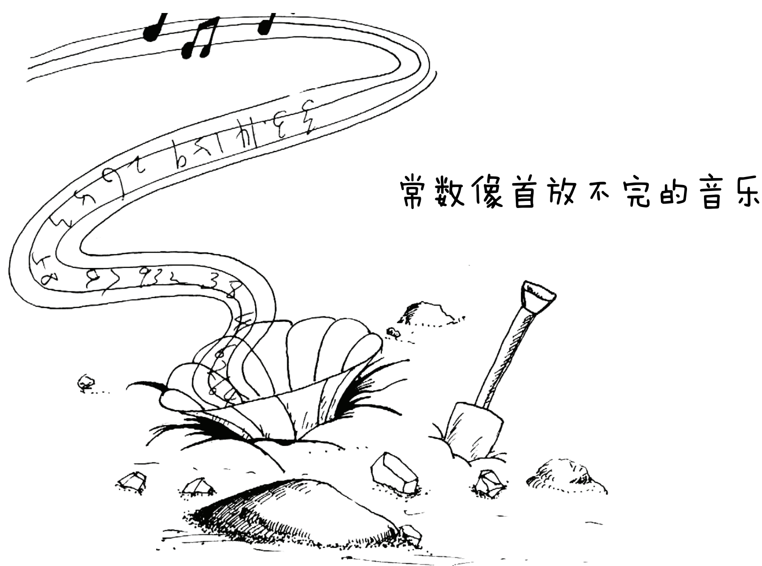
出现次数超过一半的元素

令 U 是一个有限集合，S_1，S_2，…，S_n 都是 U 的非空子集，它们满足任意多个集合的并集仍然在这些集合里。证明，一定能找到某个元素，它在至少一半的集合里出现。

不可思议，即使是最基本最离散的数学研究对象——有限集合——里面，也有让人崩溃的未解问题。

1999 年，彼达斯 · 沃伊奇克（Piotr Wojcik）用一种非常巧妙的方法证明了，存在一个元素在至少 $\frac{n}{\log_2 n}$ 个集合里出现。不过，这离目标还有很大一段距离。

14. 那些神秘的数学常数



我一直觉得，数学中的各种常数是最令人敬畏的东西，它们似乎是宇宙诞生之初上帝就已经精心选择好了的。那一串无限不循环的数字往往会让人陷入一种无底洞般的沉思——为什么这串数字不是别的，偏偏就是这个样呢？除了那些众所周知的基本常数之外，还有很多非主流的数学常数，它们的存在性和无理性同样给它们赋予了浓重的神秘色彩。现在，就让我们一起来看一看，数学当中到底有哪些神秘的常数。

$\sqrt{2} \approx 1.414\ 213\ 562\ 373\ 095\ 048\ 8$

古希腊的大哲学家毕达哥拉斯（Pythagoras）很早就注意到了数学与大千世界的联系，对数学科学的发展有着功不可没的贡献。他还创立了在古希腊影响最深远的学



派之一——毕达哥拉斯学派。毕达哥拉斯学派对数字的认识达到了审美的高度。他们相信，在这个世界中“万物皆数”，所有事物都可以用整数或者整数之比来描述。

然而，毕达哥拉斯学派的一位叫做希帕索斯（Hippasus）的学者却发现，边长为1的正方形，对角线的长度不能用整数之比来表示。这一发现无疑触犯了学派的信条，因此希帕索斯的命运非常悲惨，最后被溺死在了大海之中。与此有关的历史记载非常模糊，因此后人开始添油加醋，演绎出了这段故事的诸多版本，希帕索斯为何而死也是众说纷纭。不管怎样，希帕索斯都被人们当作了发现无理数的第一人。

利用勾股定理可知，边长为1的正方形，对角线的长度就是方程 $x^2=2$ 的唯一正数解，我们通常把它记作 $\sqrt{2}$。$\sqrt{2}$ 可能是最具代表性的无理数了，证明它的无理性有很多种方法。最常见的一种就是下面这个反证法：假设 $\sqrt{2}$ 可以表示成 $\frac{q}{p}$，并且假设 $\frac{q}{p}$ 已经是一个最简分数了。那么 $\left(\frac{q}{p}\right)^2=2$，也即 $q^2=2p^2$。这说明 q^2 是个偶数。但只有偶数的平方才能等于偶数，因此 q 一定是偶数。q 是偶数就说明 q^2 能被4整除，等式两边约掉一个2，可见 p^2 也是偶数，从而 p 是偶数。这样，p 也是偶数，q 也是偶数，那么 p 和 q 就还可以继续约分，与我们的假设矛盾。

证明还可以更简单一些。同样假设 $\frac{q}{p}$ 已经是最简分数了，那么 $\left(\frac{q}{p}\right)^2=2$，也就是 $q^2=2p^2$。注意到等式的左边是一个平方数，它只能以0、1、4、5、6结尾；等式的右边是一个平方数的两倍，它的末位则只可能是0、2、8。然而 q^2 和 $2p^2$ 是相等的，因此它们必须都以0结尾。这说明，p^2 和 q^2 里一定都含有因子5，从而 p 和 q 本身也都含有因子5，这说明 $\frac{q}{p}$ 可以继续约分，与假设矛盾。

我们还有一些更帅的方法来证明，$q^2=2p^2$ 没有正整数解。比方说，注意到，如果对一个平方数分解质因数，它必然有偶数个质因数（x^2 的所有质因数就是把 x 的质因数复制成两份）。于是，q^2 有偶数个质因数，p^2 也有偶数个质因数，$2p^2$ 就有奇数个质因数。等号左边的数有偶数个质因数，等号右边的数有奇数个质因数，这显然是不可能的，因为同一个数只有一种分解质因数的方法[①]。

① 这并不是显然成立的，它是一个需要严格证明的定理。这叫做“算术基本定理”，有时也叫做“唯一分解定理”。

无理数的出现推翻了古希腊数学体系中的一个最基本的假设，冲击了古希腊哲学中离散的世界观，引发了数学史上的第一次数学危机。

无理数虽说“无理”，但在生产生活中的用途却相当广泛。量一量你手边的书本杂志的长与宽，你会发现它们的比值都约为 1.414。这是因为通常印刷用的纸张都满足这么一个性质：把两条较短边对折到一起，得到一个新的矩形，则新矩形的长宽之比和原来一样。因此，如果原来的长宽比为 $x:1$，新的长宽比就是 $1:\frac{x}{2}$。解方程 $x:1=1:\frac{x}{2}$ 就能得到 $x=\sqrt{2}$。

圆周率 π≈3.141 592 653 589 793 238 5

不管圆有多大，它的周长与直径的比值总是一个固定的数。我们就把这个数叫做圆周率，用希腊字母 π 来表示。人们很早就认识到了圆周率的存在，对圆周率的研究甚至可以追溯到公元前。从那以后，人类对圆周率的探索就从未停止过。几千年过去了，人类对圆周率的了解越来越多，但却一直被圆周率是否有理的问题所困扰。直到 1761 年，德国数学家朗伯（Lambert）才证明了 π 是无理数。

π 是数学中最基本、最重要、最神奇的常数，它常常出现在一些与几何毫无关系的场合中。例如，全体正整数的平方的倒数和就会收敛到一个与 π 有关的数值：

$$\frac{1}{1^2}+\frac{1}{2^2}+\frac{1}{3^2}+\cdots=\frac{\pi^2}{6}$$

而任意取出两个正整数，则它们互质（最大公约数为 1）的概率为 $\frac{6}{\pi^2}$，恰好是上面这个算式答案的倒数。

自然底数 e≈2.718 281 828 459 045 235 4

在 17 世纪末，瑞士数学家伯努利（Bernoulli）注意到了一个有趣的现象：当 x 越大时，$\left(1+\frac{1}{x}\right)^x$ 将会越接近某个固定的数：

$$\left(1+\frac{1}{100}\right)^{100}\approx 2.704\,81$$



$$\left(1+\frac{1}{1000}\right)^{1000}\approx 2.716\,92$$

$$\left(1+\frac{1}{10000}\right)^{10\,000}\approx 2.718\,15$$

18 世纪的大数学家欧拉（Euler）仔细研究了这个问题，并第一次用字母 e 来表示当 x 无穷大时 $\left(1+\frac{1}{x}\right)^{x}$ 的值。他不但求出了 $\mathrm{e}\approx 2.718$，还证明了 e 是无理数。e 的用途也十分广泛，很多公式里都有 e 的身影。比方说，如果把前 n 个正整数的乘积记作 $n!$，则有斯特林（Stirling）近似公式 $n!\approx\sqrt{2\pi n}\cdot\left(\frac{n}{\mathrm{e}}\right)^{n}$。在微积分中，无理数 e 更是大显神通，$\mathrm{e}^{x}$ 的导数竟然是它本身，这使得 e 也成为了高等数学中最重要的无理数之一。

在数学中还有一个奇妙的常数 i，它叫做“虚数单位”，简单地说也就是 $\sqrt{-1}$ 的意思。虽然 $\sqrt{-1}$ 看上去非常不合理，但若承认它的存在，所有的 n 次多项式都会有恰好 n 个根（包括重根），数系瞬间变得如同水晶球一般完美。可以说，圆周率 π、自然底数 e 和虚数单位 i 是数学中最基本的三个常数。有一个等式用加法、乘法、乘方这三种最基础的运算，把这三个最基本的常数以及两个最基本的数字（0 和 1）联系在了一起，没有任何杂质，没有任何冗余，漂亮到了神圣的地步：

$$\mathrm{e}^{\pi\mathrm{i}}+1=0$$

这个等式也是由欧拉发现的，它叫做“欧拉恒等式”。《数学情报》（*The Mathematical Intelligencer*）杂志曾举办过一次读者投票活动，欧拉恒等式被评选为“史上最美的公式”。

欧拉常数 $\gamma\approx 0.577\,215\,664\,901\,532\,860\,6$

第一次看到调和级数 $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots$，很多人都以为它会收敛到一个固定的值。其实，这个级数是发散的，无限地加下去，和也将会变得无穷大。我们很容易证明这一点：把 $\frac{1}{3}$ 和 $\frac{1}{4}$ 都缩小到 $\frac{1}{4}$，把 $\frac{1}{5}$ 到 $\frac{1}{8}$ 这 4 个数都缩小到 $\frac{1}{8}$，把接下来的 8 个数都缩小到 $\frac{1}{16}$，等等，可以看出数列仍然是发散的——因为这相当于有无穷多个 $\frac{1}{2}$ 在相加。



因此，我们不但证明了 $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots$ 的发散性，还证明了数列的前 n 项之和一定大于 $\frac{1}{2}\cdot\log_2 n$ 。

虽然调和级数是发散的，但它发散的速度非常慢。把 $\frac{1}{2}$ 和 $\frac{1}{3}$ 都放大到 $\frac{1}{2}$ ，把 $\frac{1}{4}$ 到 $\frac{1}{7}$ 这 4 个数都放大到 $\frac{1}{4}$ ，把接下来的 8 个数都放大到 $\frac{1}{8}$ ，等等，可见前 n 项之和不会超过 $\log_2 n$ 个 1 相加。按此估算，数列的前 1 000 000 项之和也不到 20。

注意，$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots$ 的前 n 项之和夹在了 $\frac{1}{2}\cdot\log_2 n$ 和 $\log_2 n$ 之间，这表明它一定是对数级增加的。随着 n 的增加，$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n}$ 将会越来越接近于 $\ln n$ 。1735 年，欧拉首次发现，当 n 增加到无穷大时，$1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\cdots+\frac{1}{n}$ 和 $\ln n$ 之间的差将收敛于一个固定的值。这个值就被命名为欧拉常数，用希腊字母 γ 来表示，它约等于 0.5772。

有趣的是，虽然大家都认为欧拉常数一定是无理数，但到目前为止还没有人能够证明这一点。现在已经知道，如果欧拉常数是有理数的话，它的分母至少是 $10^{242\,080}$ 。

黄金分割 $\varphi=\frac{1+\sqrt{5}}{2}\approx1.618\,033\,988\,749\,894\,848\,2$

把一条线段分成两段，分割点在什么位置时最为美观？分在中点处，似乎太对称了不好看；分在三等分点处，似乎又显得有些偏了。人们公认，最完美的分割点应该满足这样一种性质：较长段与较短段的长度比，正好等于整条线段与较长段的长度比。这个比值就叫做黄金分割，用希腊字母 φ 来表示。若令线段的较短段的长度为 1，则 φ 就满足方程 $\varphi=\frac{1+\varphi}{\varphi}$ ，可解出 $\varphi=\frac{1+\sqrt{5}}{2}$ 。

在美学中，黄金分割有着不可估量的意义。在那些最伟大的美术作品中，每个细节的构图都充分展示了黄金分割之美。在人体中，黄金分割也无处不在——肘关节就是整只手臂的黄金分割点，膝关节就是整条腿的黄金分割点，而肚脐则位于整个人体的黄金分割点处。



在数学中，黄金分割φ也展示出了它的无穷魅力。例如，在图 1 所示的正五角星中，同一条线上三个点 A、B、C 就满足 $AB:BC=\varphi$ 。在第 12 节讲到的 8 个算术游戏中，φ也出现在了一个出人意料的地方。

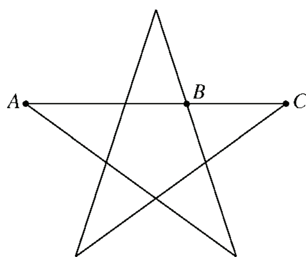


图 1

辛钦常数 $K \approx 2.685\,452\,001\ 065\,306\,445\,3$

每个实数都能写成 $a_0+\cfrac{1}{a_1+\cfrac{1}{a_2+\cdots}}$ 的形式，其中 a_0，a_1，a_2，⋯都是整数。我们就把 $[a_0;a_1,a_2,a_3,\cdots]$ 叫做该数的连分数展开。比方说，π是一个比 3 多一点点的数，大概比 3 多$\frac{1}{7}$吧。但是，这个分母 7 还不够准确。事实上π是一个小于$3+\frac{1}{7}$但是大于$3+\frac{1}{8}$的数，也就是说刚才那个分母应该比 7 要大一点点，因此π可以表示成$3+\cfrac{1}{7+\cdots}$。继续计算我们还能得出更具体的结果，π约为$3+\cfrac{1}{7+\cfrac{1}{15}}$，但是那个分母 15 比精确值还稍微小了一些，因此π可以写作$3+\cfrac{1}{7+\cfrac{1}{15+\cdots}}$。省略的部分又可以写成多少多少分之一的形式，其中分母又可以拆成一个整数部分加上一个小数部分。不断这样做下去，我们就得到了π的连分数展开：$[3;7,15,1,292,1,\cdots]$。

和小数展开比起来，连分数展开具有更加优雅漂亮的性质，这使得连分数成为了数学研究中的必修课。



在 1964 年出版的一本连分数数学课本中，数学家辛钦（Khinchin）证明了这样一个惊人的结论：除了有理数和二次整系数方程的根等特殊情况以外，几乎所有实数的连分数展开序列的几何平均数都收敛到一个相同的数，它约为 2.685 452。例如，圆周率 π 的连分数展开序列中，前 20 个数的几何平均数约为 2.628 19，前 100 个数的几何平均数则为 2.694 05，而前 1 000 000 个数的几何平均数则为 2.684 47。

目前，人们对这个神秘常数的了解并不太多。虽然辛钦常数很可能是无理数，但这一点至今仍未被证明。而辛钦常数的精确值也并不容易求出。1997 年，戴维·贝利（David Bailey）等人对一个收敛极快的数列进行了优化，但也只求出了辛钦常数的小数点后 7350 位。

康威常数 $\lambda \approx 1.303\,577\,269\,034\,296\,391\,3$

你能找出下面这个数列的规律吗？

1,

11,

21,

1211,

111221,

312211,

13112221,

1113213211,

…

这个数列的规律简单而又有趣。数列中的第一个数是 1。从第二个数开始，每个数都是对前一个数的描述：第二个数 11 就表示它的前一个数是“1 个 1”，第三个数 21 就表示它的前一个数是“2 个 1”，第四个数 1211 就表示它的前一个数是“1 个 2，1 个 1”……这个有趣的数列就叫做“外观数列”（look-and-say sequence）。

外观数列有很多有趣的性质。例如，数列中的数虽然会越来越长，但数字 4 永远不会出现。1987 年，约翰·康威发现，在这个数列中，相邻两数的长度之比越来越接



近一个固定的数。最终，数列的长度增长率将稳定在一个约为 1.303 577 的常数上。康威把这个常数命名为康威常数，并用希腊字母 λ 表示。康威证明了 λ 是无理数，它是某个 71 次方程的唯一实数解。

钱珀瑙恩常数 $C_{10} \approx 0.123\,456\,789\,101\,112\,131\,4$

把全体正整数从小到大依次写成一排，并在最前面加上小数点，便得到了一个无限小数 0.1234567891011121314…。这个数是由英国统计学家钱珀瑙恩（Champernowne）于 1933 年构造出来的，他把它命名为钱珀瑙恩常数，用符号 C_{10} 表示。与其他的数学常数相比，钱珀瑙恩常数有一个很大的不同之处：这个数仅仅是为了论证一些数学问题而人为定义出来的，它并未描述任何一个数学对象。

钱珀瑙恩常数有很多难能可贵的性质。首先，容易看出它是一个无限不循环小数，因此它也就是一个无理数。其次，它还是一个“超越数”，意即它不是任何一个整系数多项式方程的解。它还是一个“正规数”，意即每一种数字或者数字组合出现的机会都是均等的。在众多数学领域中，钱珀瑙恩常数都表现出了其非凡的意义。

15. 奇妙的心电图数列

发现数学结论的过程，无疑比数学结论本身更美妙。当你见到一个全新的几何构造，一个全新的运算法则，或者一个全新的函数定义时，不妨深入研究下去，几乎总会有惊喜发生。在这一节中，我们将从一个简单的数列出发，挖掘出越来越多的定理和猜想，体验数学发现的乐趣。

心电图数列（EKG Sequence）的定义简单而有趣：第一项为 1，第二项为 2，以后的每一项都是最小的和前一项不互质并且不曾出现过的数。换句话说，数列 $a(1)=1$，$a(2)=2$，且当 $n>2$ 时取 $a(n)$ 为所有满足以下两个条件的数中最小的那一个：该数与 $a(n-1)$ 有大于 1 的公因数，并且该数与前面 $n-1$ 项都不相等。心电图数列的前面 20 项为

$$1, 2, 4, 6, 3, 9, 12, 8, 10, 5, 15, 18, 14, 7, 21, 24, 16, 20, 22, 11, \cdots$$

为什么把它叫做心电图数列呢？原因很简单——因为把它描绘在图像上时，看上去像一张心电图（见图 1）。

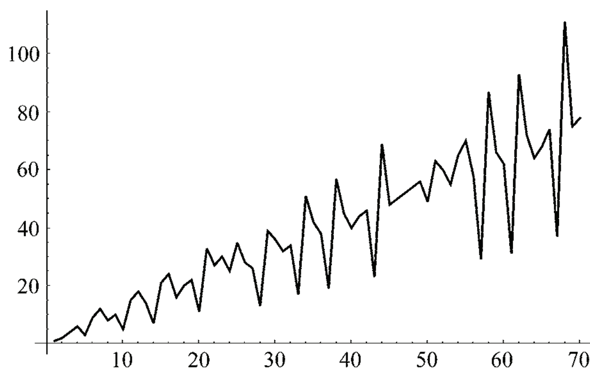


图　1



心电图数列有很多有趣的性质。例如，考虑某个质数 p ，假设数列中第一个含有质因数 p 的数是 $t\cdot p$ 。根据定义，$t\cdot p$ 和它的前一项有一个公因数。显然这个公因数不可能是 p ，因为 $t\cdot p$ 才是质因数 p 在数列中首次出现的地方；因而，这个公因数只能是 t 或者 t 的因数。由于 $t\cdot p$ 满足最小性，因此我们可以进一步得出，t 是 $t\cdot p$ 前一项的最小质因数。我们还可以推算出 $t\cdot p$ 的后一项。$t\cdot p$ 的后一项要么就是 p ，要么就是某个比 p 小的 t 的倍数。但后者是不可能的，如果存在某个 t 的倍数比 p 小而之前又没出现过，那 $t\cdot p$ 这一项本身就不会是 $t\cdot p$ 了，而将由这个 t 的倍数取代。因此，$t\cdot p$ 的后一项一定是 p 。我们还可以看出，只要 $t\neq 2$ ，这个 p 的后一项就一定是 $2p$ ；而当 $t=2$ 时，p 的后一项就只能是 $3p$ 了。也就是说，如果数列中出现了一个质数 p ，那么 $2p$ 不是它的前一项就一定是它的后一项。

有意思的是，除了 $p=2$ 以外，目前我们还没有找到 $2p$ 出现在 p 后面的情况。换句话说，人们发现，对于数列中的每个奇质数 p ，它的前一项无一例外地都是 $2p$ ，并且后面总是跟着 $3p$ 。证明或推翻这个猜想并不容易，直到最近几年才出现有关它的证明。很大程度上来说，这是整个数列呈心电图模样的最关键原因。

心电图数列有一个很漂亮的数学事实：所有的自然数都出现在了这个数列中。由这个数列的定义，每个数最多也只能出现一次。因此，心电图数列是全体自然数的一个排列。这个结论的证明堪称经典。首先我们证明引理 1：如果数列中有无穷多项都是某个质数 p 的倍数，那么 p 的任意一个倍数都出现在了数列中。证明的基本思路是反证。无妨假定 $k\cdot p$ 是最小的不在数列中的 p 的倍数，那么我们总能找到一个充分大的 N ，使得从第 N 项开始所有数都不小于 $k\cdot p$ 。然而数列中有无穷多项都是 p 的倍数，因此在第 N 项后面一定能找到一个 p 的倍数，这个数的下一项就只可能是 $k\cdot p$ 了，矛盾。

我们可以故技重施，继续证明引理 2：如果某个质数 p 的任意一个倍数都出现在了数列中，那么所有正整数都出现在了数列中。反证，假设 k 是最小的不在数列中的数，我们总能找到一个充分大的 N ，使得从第 N 项起后面的所有数都不小于 k 。由于质数 p 的任一倍数都在数列里，因此 $k\cdot p$ 的任一倍数都在数列里，即数列中有无穷多项都是 k 的倍数。那么，第 N 项之后一定存在一个 k 的倍数，它的下一项就只可能是 k 了，矛盾。



接下来就是最妙的地方了。我们可以利用上面两个引理立即得知，所有正整数都出现在了数列中。假设数列中所有项的所有质因数只有有限多种，由于整个数列有无穷多个数，因此至少有一种质因数出现了无穷多次，由引理 1 可知这个质因数的所有倍数都在数列里，由引理 2 知所有正整数都出现在了数列中，与“质因数只有有限多种”的假设矛盾。因此，数列中包含有无穷多种质因数。而前面说过，数列中第一个含有质因数 p 的项，其下一项一定是质数 p 。因此，数列中出现了无穷多个质数。而质数 p 的前一项或者后一项必有一个是 $2p$ ，因此质因数 2 出现了无数多次。由引理 1 可知 2 的所有倍数都在数列里。由引理 2 可知所有正整数都在数列中了。

心电图数列还有很多优美的性质和尚未解决的猜想。如图 2 所示，把前面 500 多个数描绘在图像上，容易看出整个图像大致成三条斜线，其中两条稀疏的线明显是由形如 p 和 $3p$ 的数组成。于是有人猜想，如果把所有 p 和 $3p$ 都变成 $2p$ ，整个数列在渐近意义上与 $f(n)=n$ 等价。

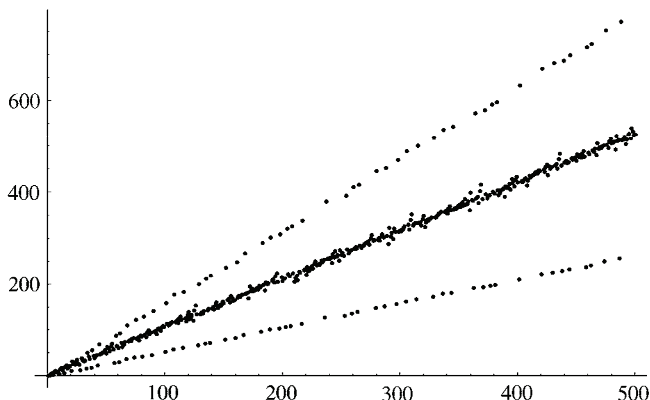


图 2

由此我们又想到一个问题，既然 $a(n)$ 与 n 相差不远，那么它们之间的大小关系究竟如何？作出 $a(n)-n$ 的图像（见图 3），我们立即得出一个新的猜想：排除 $a(n)$ 为质数的情况，则几乎所有 $a(n)$ 都大于 n 。



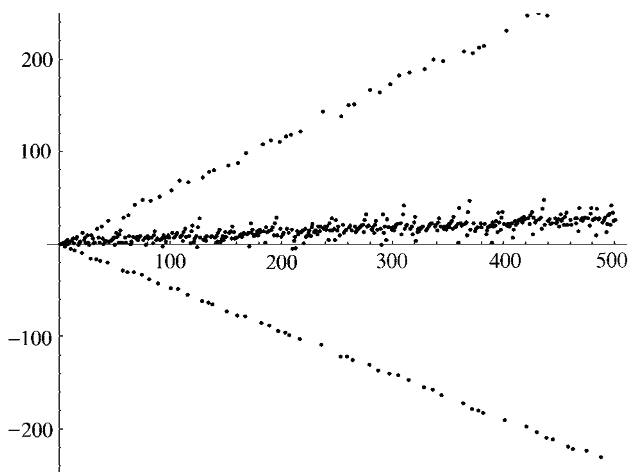


图 3

根据已有资料，在这两个问题中，前一个问题好像已经得到了证明，后一个问题则是最近才提出的猜想，还有待人们继续探索。

16. 不可思议的分形图形

讲数学之美，分形图形是不可不讲的。如果说有什么东西能够让数学和艺术直接联系在一起，答案毫无疑问就是分形图形。

让我们先来看一个简单的例子。首先画一个线段，然后把它平分成三段，去掉中间那一段并用两条等长的线段代替。这样，原来的一条线段就变成了四条小的线段。用相同的方法把每一条小的线段的中间三分之一替换成一座小山，得到了 16 条更小的线段。然后继续对这 16 条线段进行类似的操作，并无限地迭代下去。图 1 是这个图形前五次迭代的过程，可以看到第五次迭代后图形已经相当复杂，我们已经无法看清它的全部细节了。

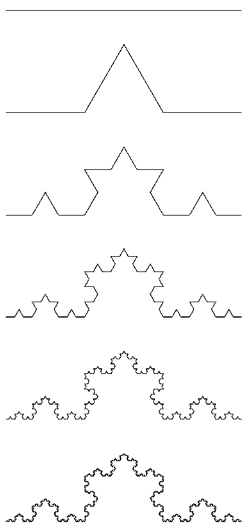


图　1



你可能注意到一个有趣的事实：整个线条的长度每一次都变成了原来的$\frac{4}{3}$。如果最初的线段长度为一个单位，那么第一次操作后总长度变成了$\frac{4}{3}$，第二次操作后总长度增加到$\frac{16}{9}$，第n次操作后总长度为$\left(\frac{4}{3}\right)^n$。毫无疑问，操作无限进行下去，这条曲线将达到无限长。难以置信的是这条无限长的曲线却“始终只有那么大”。

现在，我们像图 2 那样，把 3 条这样的曲线首尾相接组成一个封闭图形。这时，有趣的事情发生了，这个雪花状的图形有着无限长的边界，但是它的总面积却是有限的。有人可能会说，为什么面积是有限的呢？虽然从图 2 看结论很显然，但这里我们还是要给出一个简单的证明。3 条曲线中每一条在第n次迭代前都有4^{n-1}条长为$\left(\frac{1}{3}\right)^{n-1}$的线段，迭代后多出的面积为$4^{n-1}$个边长为$\left(\frac{1}{3}\right)^n$的等边三角形。把$4^{n-1}$扩大到$4^n$，再把所有边长为$\left(\frac{1}{3}\right)^n$的等边三角形扩大为同样边长的正方形，总面积仍是有限的，因为无穷级数$\frac{4}{9}+\frac{4^2}{9^2}+\frac{4^3}{9^3}+\cdots$是收敛的。很难相信，这一块有限的面积，竟然是用无限长的曲线围成的。

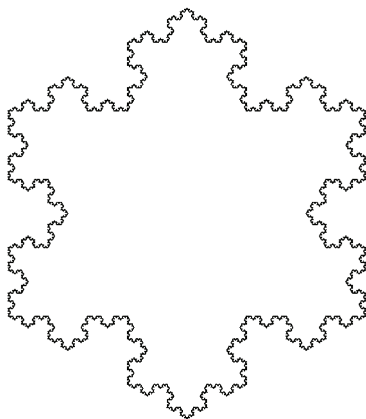


图 2

这让我们开始质疑“周长”的概念了：剪下一个直径为 1 厘米的圆形纸片，它的周长真的就是π厘米吗？拿放大镜看看，我们就会发现纸片边缘并不是平整的，上面



充满了小锯齿。再用显微镜观察，说不定每个小锯齿上也长有很多小锯齿。然后，锯齿上有锯齿，锯齿上又有锯齿，周长永远也测不完。分形领域中有一个经典的说法，“英国的海岸线有无限长”，其实就是这个意思。

上面这个神奇的雪花图形叫做科赫雪花，那条无限长的曲线就叫做科赫曲线。他是由瑞典数学家冯·科赫（Helge von Koch）最先提出来的。

分形这一课题提出的时间比较晚。科赫曲线于 1904 年提出，是最早提出的分形图形之一。我们仔细观察一下这条特别的曲线。它有一个很强的特点：你可以把它分成若干部分，每一个部分都和原来一样（只是大小不同）。这样的图形叫做“自相似”（self-similar）图形。自相似是分形图形最主要的特征，它往往都和递归、无穷之类的东西联系在一起。比如，自相似图形往往是用递归法构造出来的，可以无限地分解下去。一条科赫曲线包含有无数大小不同的科赫曲线。你可以对这条曲线的尖端部分不断放大，但你所看到的始终和最开始一样。它的复杂性不随尺度减小而消失。另外值得一提的是，它是一条连续的，但处处不光滑（不可微）的曲线。曲线上的任何一个点都是尖点。

分形图形有一种特殊的计算维度的方法。我们可以看到，在有限空间内就可以达到无限长的分形曲线似乎已经超越了一维的境界，但说它是二维图形又还不够。1918 年，数学家费利克斯·豪斯道夫（Felix Hausdorff）提出了豪斯道夫维度，它就是专门用来对付这种情况的。简单地说，豪斯道夫维度描述了对分形图形进行缩放后，图形所占空间大小的变化与相似比的关系。例如，把正方形的边长扩大到原来的 2 倍后，正方形的面积就将变成原来的 4 倍；若把正方形的边长扩大到原来的 3 倍，则其面积就将变成原来的 9 倍。事实上，两个正方形的相似比为 $1:a$，它们的面积比就应该是 $1:a^2$，那个指数 2 就是正方形的豪斯道夫维度。类似地，两个立方体的相似比为 $1:a$，它们的体积比就是 $1:a^3$，这就告诉了我们，立方体的豪斯道夫维度是 3。然而，一条大科赫曲线包含了 4 条小科赫曲线，但大小科赫曲线的相似比却只有 1:3。也就是说，把小科赫曲线放大到原来的 3 倍，所占空间会变成原来的 4 倍！因此科赫曲线的豪斯道夫维度为 $\log_3 4$。它约等于 1.26，是一个介于 1 和 2 之间的实数。直观地说，科赫曲线既是曲线，又非曲线，它介于线与面之间。

很多分形图形的维度都介于 1 和 2 之间。比如说谢尔宾斯基（Sierpinski）三角形：



像图 3 那样，把一个三角形分成 4 等份，挖掉中间那一份，然后继续对另外 3 个三角形进行这样的操作，并且无限地递归下去。每一次迭代后整个图形的面积都会减小到原来的 $\frac{3}{4}$，因此最终得到的图形面积显然为 0。因而和科赫曲线正好相反，它已经不能算二维图形了，但说它是一维的似乎也有些过了。事实上，它的豪斯道夫维度是 $\log_2 3$，也是一个介于 1 和 2 之间的图形。

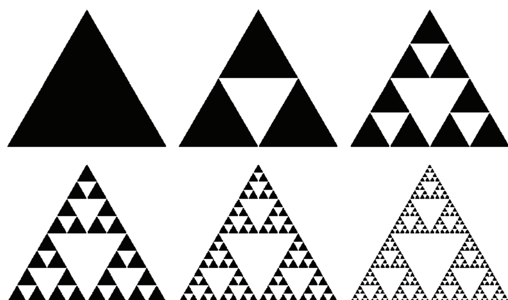


图 3

谢尔宾斯基三角形的另一种构造方法如图 4 所示。把正方形分成四等份，去掉右下角的那一份，并且对另外 3 个正方形递归地操作下去。挖几次后把脑袋一歪，你就可以看到一个等腰直角的谢尔宾斯基三角形了。



图 4

谢尔宾斯基三角形还有一些非递归的构造。1983 年，斯蒂芬 · 沃尔夫勒姆（Stephen Wolfram）发现，在一个网格中，从一个黑色格子开始，不断按规则生成下一行的图形（见图 5），也能得到谢尔宾斯基三角形。这种图形生成方法有一个很酷的名字，叫做“细胞自动机”。

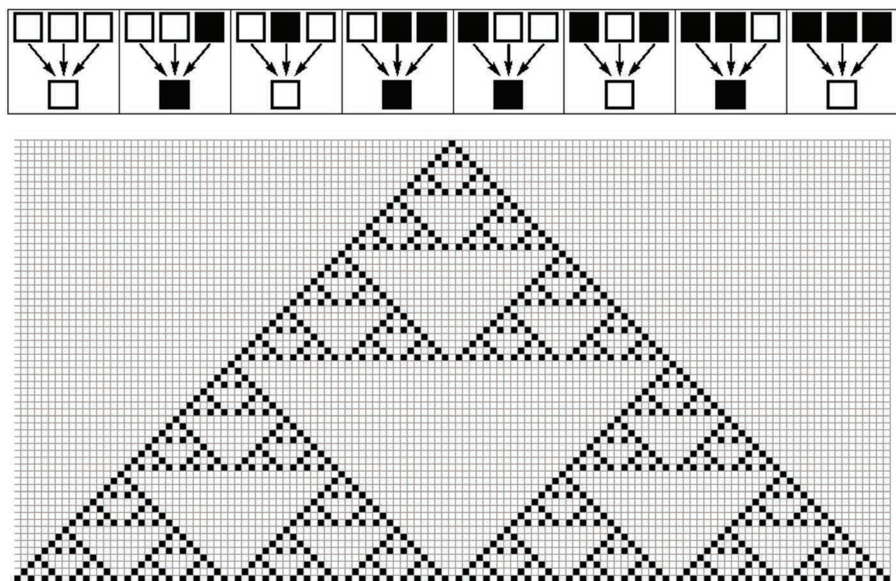


图　5

谢尔宾斯基三角形有一个神奇的性质：如果某一个位置上有点（没被挖去），那么它与原三角形顶点的连线上的中点处也有点。这给出了一个更为诡异的谢尔宾斯基三角形构造方法：给出三角形的 3 个顶点，然后从其中一个顶点出发，每次随机向任意一个顶点移动 $\frac{1}{2}$ 的距离（走到与那个顶点的连线的中点上），并在该位置作一个标记；无限次操作后所有的标记就组成了谢尔宾斯基三角形。

杨辉三角与谢尔宾斯基三角形之间也有不可思议的关系。如图 6，把杨辉三角中的奇数和偶数用不同的颜色区别开来，你会发现由此得到的正是谢尔宾斯基三角形。也就是说，二项式系数（或者说组合数）的奇偶性竟然可以表现为一个分形图形！这相当于给出了谢尔宾斯基三角形的第五种构造方法。利用简单的代数方法生成如此优雅的图形，实在是令人叹为观止。请记住谢尔宾斯基三角形这个最经典的分形图形，因为在未来的某个时刻，我们将会在某个出人意料的地方用到它。

大家或许已经看到了数学的奇妙之处：一个如此简单的公式，竟能形成如此美观精细的图形。说到这里，我们不得不提另一个奇迹般的分形图形。



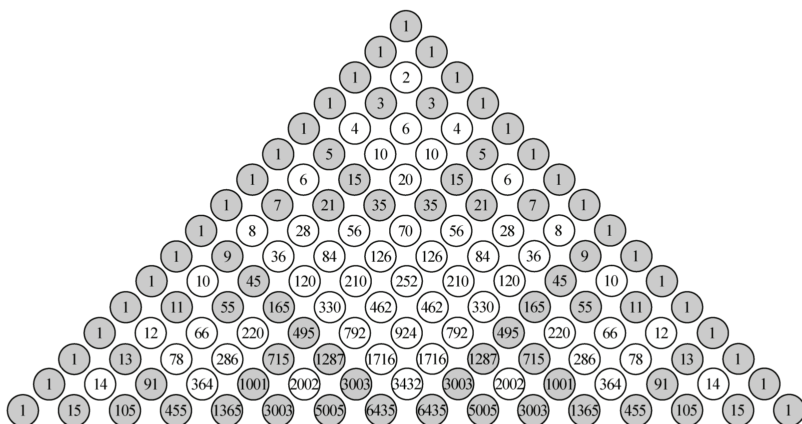


图 6

考虑函数 $f(z)=z^2-0.75$。固定 z_0 的值后，我们可以通过不断地迭代算出一系列的 z 值：$z_1=f(z_0)$，$z_2=f(z_1)$，$z_3=f(z_2)$，…。比如，当 $z_0=1$ 时，我们可以依次迭代出：

$$z_1=f(1.0)=1.0^2-0.75=0.25$$

$$z_2=f(0.25)=0.25^2-0.75=-0.6875$$

$$z_3=f(-0.6875)=(-0.6875)^2-0.75=-0.2773$$

$$z_4=f(-0.2773)=(-0.2773)^2-0.75=-0.6731$$

$$z_5=f(-0.6731)=(-0.6731)^2-0.75=-0.2970$$

……

可以看出，z 值始终在某一范围内，并将最终收敛到某一个值上。

但当 $z_0=2$ 时，情况就不一样了。几次迭代后我们将立即发现 z 值最终会趋于无穷大：

$$z_1=f(2.0)=2.0^2-0.75=3.25$$

$$z_2=f(3.25)=3.25^2-0.75=9.8125$$

$$z_3=f(9.8125)=9.8125^2-0.75=95.535$$

$$z_4=f(95.535)=95.535^2-0.75=9126.2$$

$$z_5=f(9126.2)=9126.2^2-0.75=83\,287\,819.2$$

……

经过计算，我们可以得到如下结论：当 z_0 属于 $[-1.5,1.5]$ 时，z 值始终不会超出某个范围；而当 z_0 小于 -1.5 或大于 1.5 后，z 值最终将趋于无穷。

现在，我们把这个函数扩展到整个复数范围。对于复数 $z_0 = a + b\text{i}$，取不同的 a 值和 b 值，函数迭代的结果不一样：对于有些 z_0，函数值始终约束在某一范围内；而对于另一些 z_0，函数值则将发散到无穷。我们把满足前一种情况的所有初始值 z_0 所组成的集合称为朱利亚集，它是以法国数学家加斯顿·朱利亚（Gaston Julia）的名字命名的。

由于复数对应了平面上的点，因此我们可以用一个平面图形直观地展现出朱利亚集。我们用黑色表示所有属于朱利亚集的 z_0；对于其他的 z_0，我们用不同的颜色来区别不同的发散速度，颜色越浅表示发散速度越慢，颜色越深表示发散速度越快。难以置信，由此得到的图形竟然是一个看上去非常复杂的分形图形（见图 7）。

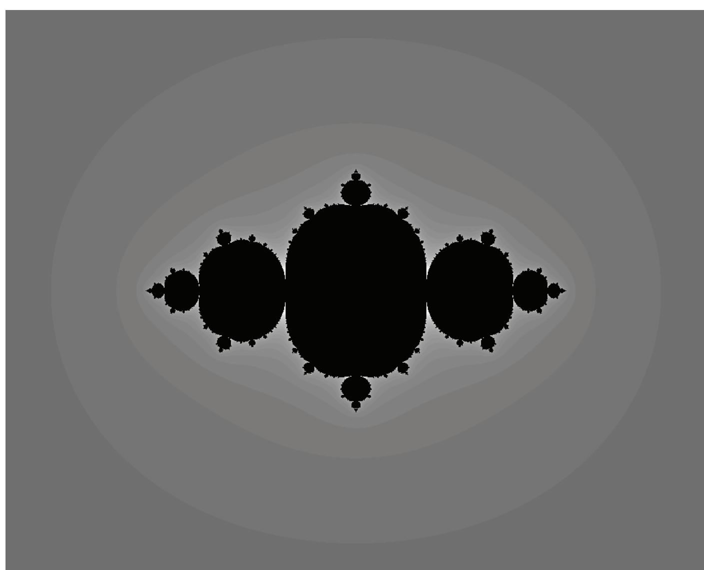


图　7

这个美丽的分形图形就是 $f(z) = z^2 - 0.75$ 时的朱利亚集。如果我们把 -0.75 换成别的数，比如 $-0.8 + 0.15\text{i}$ 呢？这将会带来另一个完全不同的分形图形，图 8 就是 $f(z) = z^2 - 0.8 + 0.15\text{i}$ 所对应的朱利亚集。





图 8

事实上，对于复数函数 $f(z)=z^2+c$ ，每取一个不同的复数 c ，我们都能得到一个不同的朱利亚集分形图形，并且令人吃惊的是，每一个分形图形都是那么美丽，其中有些经典的朱利亚集甚至有它自己的名字。图 9 就是 $c=-1.755$ 时的朱利亚集，俗称“飞机”。

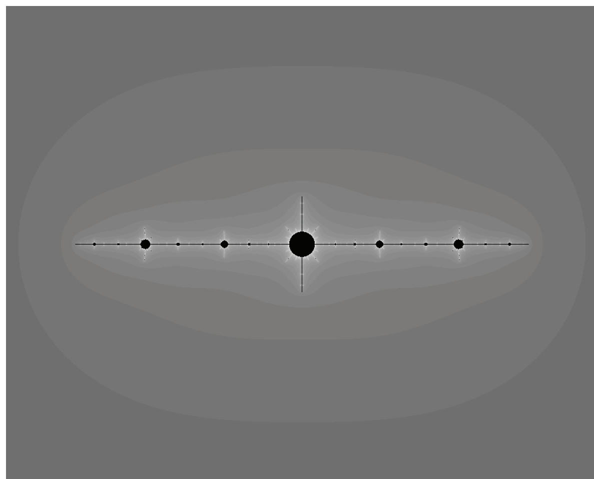


图 9

图 10 则是 $c=-0.123+0.745\mathrm{i}$ 所对应的朱利亚集。它也有一个形象的名字——杜瓦地兔子。这是以法国数学家阿德里安 · 杜瓦地（Adrien Douady）的名字命名的。

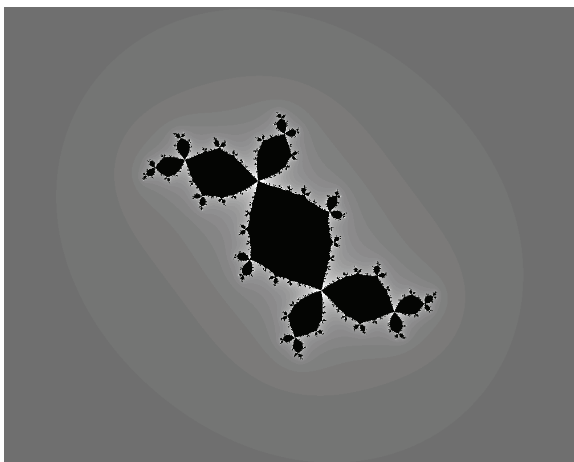


图 10

你甚至会不相信，这种简单而机械的过程可以生成如此美丽的图形。

不过，并不是所有的复数 c 都对应了一个连通的朱利亚集。图 11 所示的就是 $c = 0.3$ 时的朱利亚集。这仍然是一个漂亮的分形图，但它和前面的图像有一个很大的区别——图像里不再有连通的黑色区域了。这是因为，真正属于朱利亚集的点都是一个个离散的点（分布在图中的各个白色亮斑中），我们已经无法从图像上直接观察到了。我们能看到的，都是那些将会导致函数值发散到无穷的点，只是它们的发散速度有所不同。

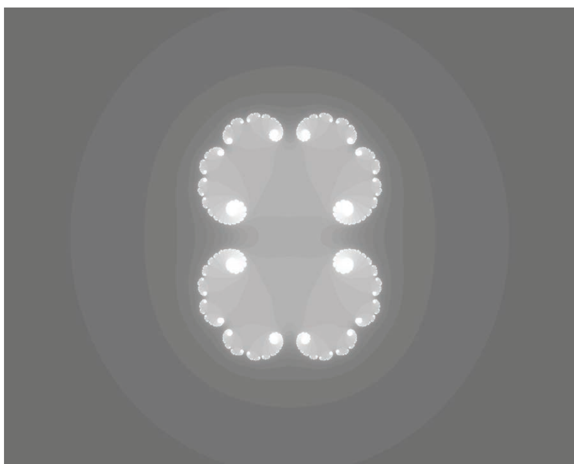


图 11



于是，我们自然想到了一个问题：哪些复数c对应着连通的朱利亚集呢？数学家贝努瓦·曼德尔布罗特（Benoit Mandelbrot）是最早对这个问题进行系统研究的人之一，因此我们通常把所有使得朱利亚集形成一块连通区域的复数c所组成的集合叫做曼德尔布罗特集。注意，曼德尔布罗特集也是一个由复数构成的集合，它也能表现在一个平面上。神奇的是，曼德尔布罗特集本身竟然又是一个漂亮的分形图形（见图 12）!

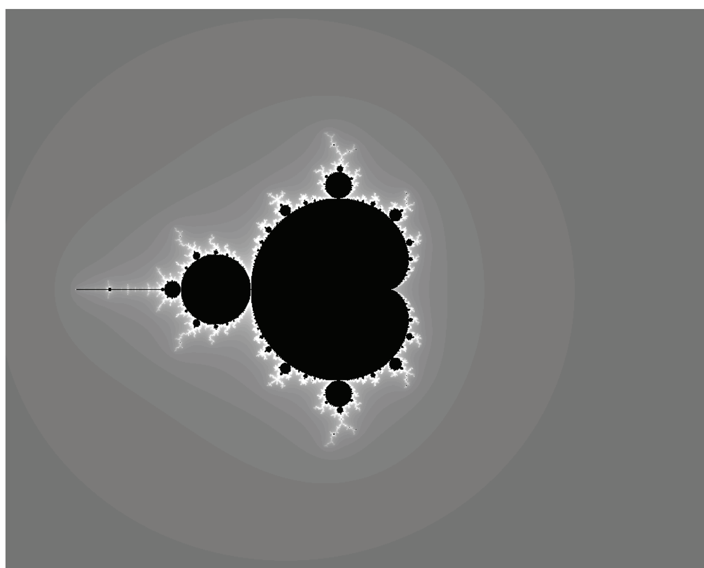


图　12

有一个重要的定理指出，一个朱利亚集是连通的，当且仅当$z_0=0$在这个朱利亚集里。换句话说，为了判断一个朱利亚集是否连通，我们只需要测试一下$z_0=0$时的迭代结果即可。因此，我们有了曼德尔布罗特集的一个等价的定义，也就是所有不会让零点发散的复数c组成的集合。图 12 其实就是依据这个原理制作的，其中黑色的区域表示曼德尔布罗特集，即那些不会让零点发散的复数c；其他的点所对应的复数c都将会让零点发散，浅色代表发散慢，深色代表发散快。

前面说过，分形图形是可以无限递归下去的，它的复杂度不随尺度减小而消失。曼德尔布罗特集中大小两个主要圆盘相接处所产生的深沟叫做“海马谷”（sea horse valley）。图 13 展示了它的一个局部大图。它的细节非常丰富，你会看到很多像海马尾巴一样的钩子以一种分形的方式排列开来。

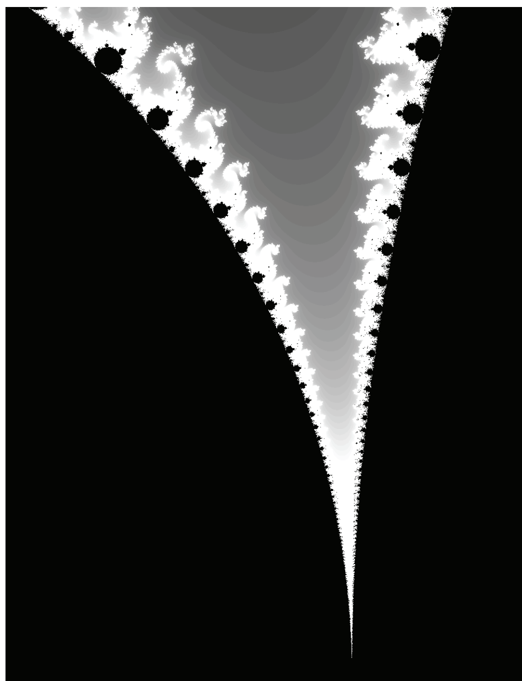


图　13

图 14 则展现了曼德尔布罗特集最右边那个深沟的景观，它也有一个名字，叫做“大象谷”（elephant valley）。

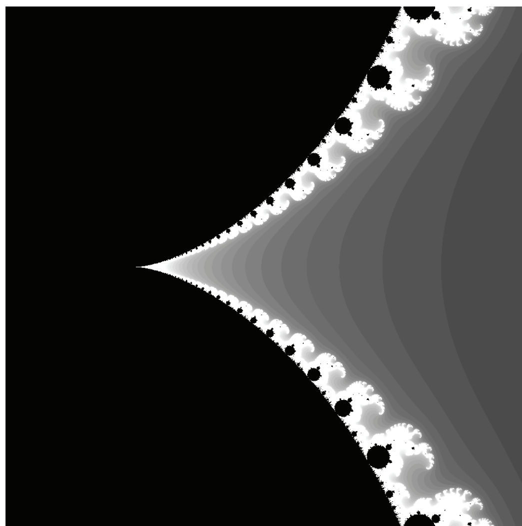


图　14



曼德尔布罗特集里值得放大的地方太多了。仔细看看曼德尔布罗特集最上方的白色触须里，是不是有一些小黑点？让我们放大一下，看看它们究竟是什么吧（见图 15）。

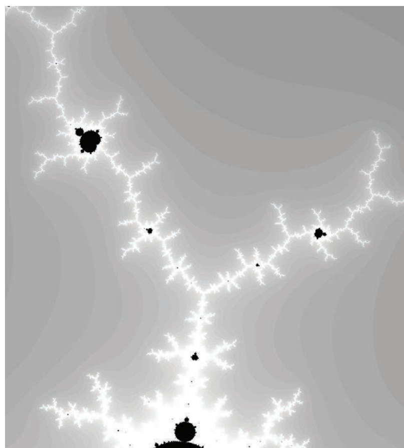


图　15

你会发现，它们竟然是曼德尔布罗特集本身的形状！此时，你应该能体会到曼德尔布罗特集的深邃与神秘了吧。

如果有人提到了数学之美，我首先想到的便是曼德尔布罗特集，简单的函数迭代竟能产生如此令人震撼的结果，壮观到了让人敬畏的地步。

如果你整天都被各种数学公式折磨，并且因此厌恶数学的话，不妨在网上找些曼德尔布罗特集的图片来看看。曼德尔布罗特集完美地诠释了我非常喜欢的一个比喻：数学不只是一堆公式，正如天文学不只是一堆望远镜（见图 16）。

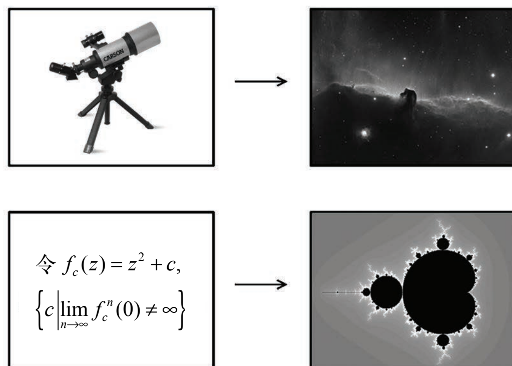


图　16



17. 几何之美：三角形的心

我曾经教过一段时间的初中数学竞赛课，下面这些内容来源于我在初三数学竞赛课的一份讲义。这节课的主题本是四点共圆，但由此引出了三角形中很多漂亮的性质，让人深感平面几何之美。不管你是否喜爱数学，你都会被这些奇妙的结论所震撼。

三角形的奇迹首先表现在各个“心”上：三角形内部的每一组有几何意义的线条都交于一点（见图 1）。3 条角平分线交于一点，这个点就叫做三角形的“内心”，它是三角形内切圆的圆心；3 条边的中垂线交于一点，这个点就叫做三角形的“外心”，它是三角形外接圆的圆心；三角形的 3 条中线也交于一点，这个点叫做三角形的“重心”，因为它真的就是这个三角形的重心。用力学方法可以很快推导出，它位于各中线的三等分点处。这些心将会在本节后面某个出人意料的地方再次出现。

三角形的 3 条高也不例外——它们也交于一点，这个点就叫做三角形的垂心。

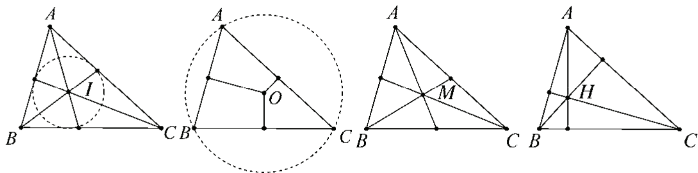


图　1

垂心看上去很不起眼，但深入研究后即会冒出很多奇妙的结论。由于两个斜边重合的直角三角形将会产生出共圆的四点，因此画出三角形的 3 条高后，会出现大量四点共圆的情况，由此将挖掘出一连串漂亮的结论。让我们先来看一个简单而直接的结论。



定理 若 D、E、F 分别是 $\triangle ABC$ 三边的高的垂足，则 $\angle 1=\angle 2$（见图 2）。

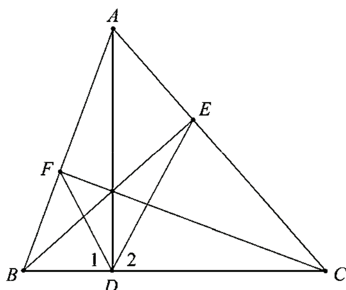


图 2

证明 由于 $\angle AFC=\angle ADC=90°$，因此 A、C、D、F 四点共圆。由于圆内接四边形对角互补，因此 $\angle 1=180°-\angle CDF=\angle A$。同理，由 A、B、D、E 四点共圆可知 $\angle 2=\angle A$。因此 $\angle 1=\angle 2$。

如果把三边垂足构成的三角形称作“垂足三角形”的话，我们就有了下面这个听上去很帅的推论。

推论 三角形的垂心是其垂足三角形的内心（见图 3）。

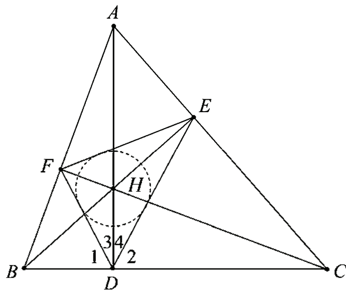


图 3

证明 因为 AD 垂直于 BC，而刚才又证明了 $\angle 1=\angle 2$，因此 $\angle 3=\angle 4$，即 HD 平分 $\angle EDF$。类似地，HE、HF 都是 $\triangle DEF$ 的内角平分线，因此 H 是 $\triangle DEF$ 的内心。

另一个有趣的推论如下。



推论　将△*ABC* 沿 *AC* 翻折到△*AB′C*，假设 *EF* 翻折到了 *EF′*，则 *EF′*和 *DE* 共线。（见图 4）。

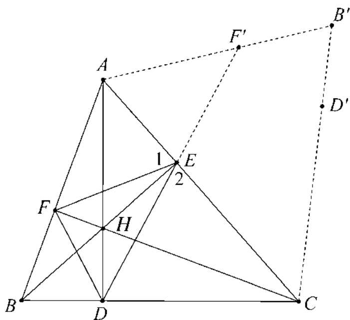


图　4

证明　这可以直接由图 4 中的∠1=∠2 推出。

1775 年，法尼亚诺（Fagnano）曾经提出了下面这个问题：在给定的锐角三角形 *ABC* 中，什么样的内接三角形具有最短的周长。这个问题就被称作“法尼亚诺问题”。法尼亚诺自己给出了答案：周长最短的内接三角形就是垂足三角形。下面我们就来证明这个结论。

定理　在△*ABC* 的所有内接三角形中，垂足三角形△*DEF* 拥有最短的周长。

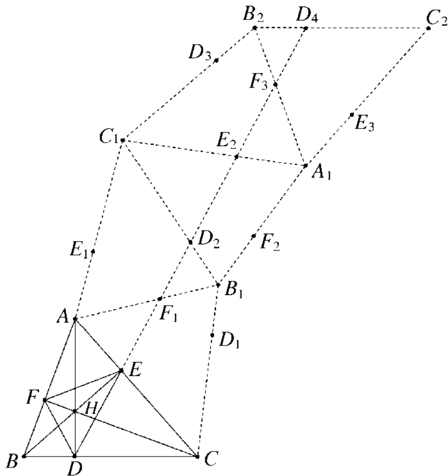


图　5



证明 像图 5 那样，把三角形翻折五次，得到折线段 $DEF_1D_2E_2F_3D_4$。这条折线段的总长度等于内接三角形 DEF 周长的两倍。注意到，由前面提到的垂足三角形的性质可知，这条折线段正好组成了一条直线段。另外，注意到如此翻折之后，BC 和 B_2C_2 是平行且相等的，而且 D 和 D_4 位于两线段上相同的位置，因此从 D 到 D_4 的折线段总长总是以直线段 DD_4 最短。这就说明了，垂足三角形 $\triangle DEF$ 拥有最短的周长。

不过，这还不够震撼，垂心还有不少的本事。四点共圆还会给我们带来其他的等角。

定理 若 D、E、F 分别是 $\triangle ABC$ 三边的高的垂足，则 $\angle 1=\angle 2$（见图 6）。

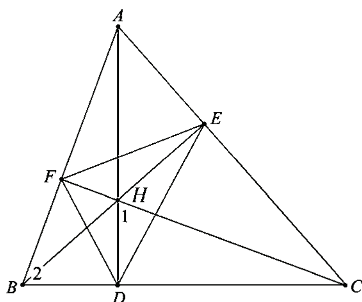


图 6

证明 由于 $\angle BFH=\angle BDH=90°$，因此 B、F、H、D 四点共圆，因此 $\angle 1=180°-\angle FHD=\angle 2$。

这将给我们带来下面这个非常漂亮的推论。

推论 把 $\triangle ABC$ 的垂心 H 沿 BC 边翻折到 H'，则 H' 在 $\triangle ABC$ 的外接圆上（见图 7）。

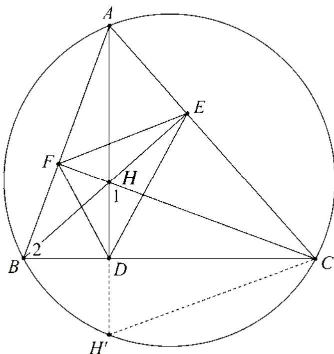


图 7



证明　由于 H 和 H'沿 BC 轴对称，因此$\angle H'=\angle 1$。而前面已经证明过了，$\angle 1=\angle 2$。因此，$\angle H'=\angle 2$。而$\angle H'$和$\angle 2$都是 AC 所对的角，它们相等就意味着 A、C、H'、B 是四点共圆的。

换一种描述方法，这个结论还可以变得更酷：

推论　把$\triangle ABC$ 的垂心 H 沿三边分别翻折到 H_1、H_2、H_3，则 A、B、C、H_1、H_2、H_3 六点共圆（见图 8）。

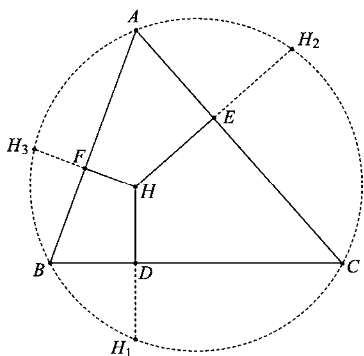


图　8

证明　这可以直接由前面的结论得到。

另一个更加对称美观的结论如下：

推论　若 D、E、F 分别是$\triangle ABC$ 三边的高的垂足，H 是垂心，则 $AH\cdot DH=BH\cdot EH=CH\cdot FH$（见图 9）。

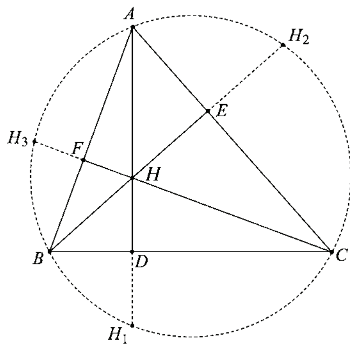


图　9



证明 做出△*ABC* 的外接圆，然后延长 *HD*、*HE*、*HF*，它们与外接圆的交点分别记作 H_1、H_2、H_3。前面的结论告诉我们，$HH_1=2HD$，$HH_2=2HE$，$HH_3=2HF$。而相交弦定理（经过圆内一定点的弦，被该点分得的两条线段的长度乘积为定值，这可以由相似三角形迅速得证）告诉我们，$AH\cdot HH_1=BH\cdot HH_2=CH\cdot HH_3$。各等量同时除以 2，就有 $AH\cdot DH=BH\cdot EH=CH\cdot FH$。

让我们再来看一个与外接圆有关的定理。

定理 若 *D*、*E*、*F* 分别是△*ABC* 三边的高的垂足，*H* 是垂心。过 *C* 作 *BC* 的垂线，与△*ABC* 的外接圆交于点 *G*。则 *CG*=*AH*(见图 10)。

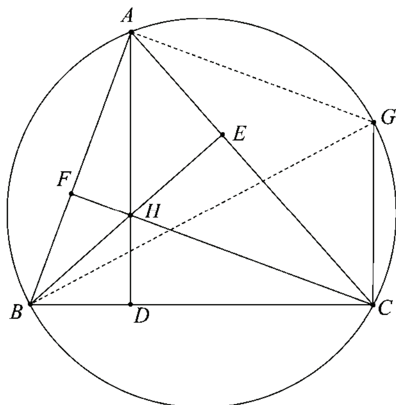


图 10

证明 我们将证明四边形 *AHCG* 的两组对边分别平行，从而说明它是一个平行四边形。注意到 *CG* 和 *AD* 都垂直于 *BC*，因此 *CG* 和 *AD* 是平行的。由于∠*BCG* 是直角，这说明 *BG* 是圆的直径，也就说明∠*BAG* 也是直角，即 *GA* 垂直于 *AB*。而 *CF* 也垂直于 *AB*，所以 *AG* 与 *CF* 平行。因而四边形 *AHCG* 是平行四边形，*CG*=*AH*。

它也能带来一个更帅的推论。

推论 若 *H* 是△*ABC* 的垂心，*O* 是△*ABC* 的外心，则 *O* 到 *BC* 的垂线段 *OM* 与 *AH* 平行，并且是 *AH* 长度的一半（见图 11）。



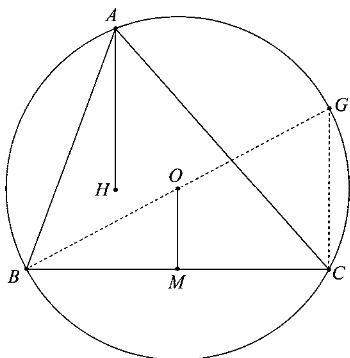


图 11

证明 前面我们证明了，图 11 中的 CG 与 AH 平行且相等。注意到 BG 是外接圆的直径，BG 的中点就是圆心，也就是 $\triangle ABC$ 的外心 O。垂线段 OM 是 $\triangle BCG$ 的中位线，它平行且等于 CG 的一半，从而也就平行且等于 AH 的一半。

好了，下面大家将会看到的就是初等几何的瑰宝。

推论 三角形的垂心、重心和外心共线，且重心在垂心和外心连线的三等分点处（见图 12）。

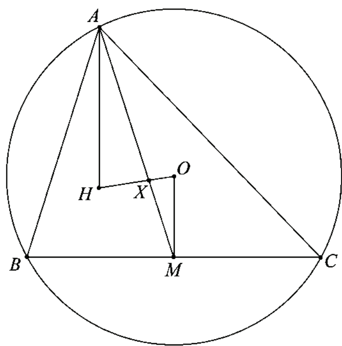


图 12

证明 把 AM 和 HO 的交点记作 X。刚才我们已经证明了，AH 与 OM 平行，且长度之比为 2:1。因此，$\triangle AHX$ 和 $\triangle MOX$ 相似，相似比为 2:1。由此可知，HX:XO=2:1，即 X 在线段 HO 的三等分点处。另外，AX:XM=2:1，也就是说 X 在三角形中线 AM 的 2:1 处。这说明，X 正是三角形的重心！



任意给定一个三角形，它的垂心、重心和外心三点共线，且重心将垂心和外心的连线分成 2:1 两段。这个美妙的结论是大数学家欧拉在 1765 年发现的，因而三角形中垂心、重心、外心所成的直线也就叫做“欧拉线”。

在三角形中，与内心、外心、重心、垂心有关的结论还有很多，我们很难在一篇文章里把它们讲完。事实上，三角形的心也不止这么几个。1994 年，美国数学教授克拉克·金伯林（Clark Kimberling）开始收集历史上被数学家们研究过的三角形的心，并建立了“三角形中心百科全书”的网站。这个网站记录了几乎所有目前已知的三角形的心。在这部百科全书里，每个三角形的心都有一个编号，编号为 n 的心就用符号 $X(n)$ 来表示，其中 $X(1)$ 到 $X(8)$ 分别为内心、重心、外心、垂心、九点圆圆心、类似重心、热尔岗（Gergonne）点和奈格尔（Nagel）点。不但每个心都有自己独特的几何性质，各个心之间还有大量共线、共圆的关系。

这个网站的地址是 http://faculty.evansville.edu/ck6/encyclopedia/ETC.html。目前，整个网站已经收集了近 5000 个三角形的心，且这个数目还在不断增加。



18. 数学之外的美丽：幸福结局问题

这是一个小故事，一个结局很幸福的小故事。

1933 年，匈牙利数学家乔治·塞凯赖什（George Szekeres）只有 22 岁。那时，他常常和朋友们在匈牙利的首都布达佩斯讨论数学。这群人里面还有同样生于匈牙利的数学怪才埃尔德什大神。不过当时，埃尔德什只有 20 岁。

在一次数学聚会上，一位叫埃丝特·克莱因（Esther Klein）的美女同学提出了这么一个结论：在平面上随便画 5 个点（其中任意三点不共线），那么一定有 4 个点，它们构成一个凸四边形。塞凯赖什和埃尔德什等人想了好一会儿，依然不知道该怎么证明这个结论。于是，美女同学得意地宣布了她的证明：如图 1，这 5 个点的凸包（覆盖整个点集的最小凸多边形）只可能是五边形、四边形和三角形。前两种情况都已经不用再讨论了，而对于第三种情况，把三角形内的两个点连成一条直线，则三角形的 3 个顶点中一定有 2 个顶点在这条直线的同一侧，这 4 个点便构成了一个凸四边形。

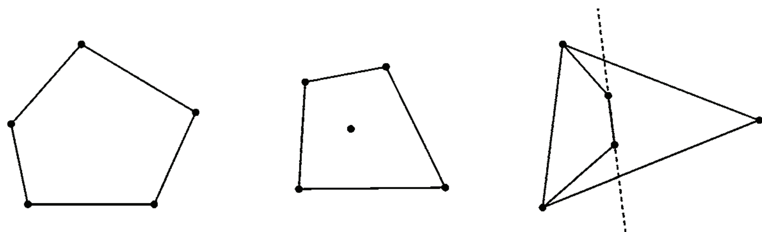


图 1

众人大呼精彩。之后，埃尔德什和塞凯赖什仍然对这个问题念念不忘，于是尝试对其进行推广。最终，他们于 1935 年发表论文，成功地证明了一个更强的结论：对于任意一个正整数 $n \geqslant 3$，总存在一个正整数 m，使得只要平面上的点有 m 个（并且任意三点不共线），那么一定能从中找到一个凸 n 边形。埃尔德什把这个问题命名为



“幸福结局问题”（Happy Ending problem），因为这个问题让塞凯赖什和美女同学克莱因走到了一起，两人在 1936 年喜结良缘。

对于一个给定的 n，不妨把需要的最少点数记作 $f(n)$。求出 $f(n)$ 的准确值是一个不小的挑战。由于平面上任意不共线三点都能确定一个三角形，因此 $f(3)=3$。克莱因的结论则可以简单地表示为 $f(4)=5$。

当 $n=5$ 时，8 个点是不够的。图 2 就是 8 个不含凸五边形的点。

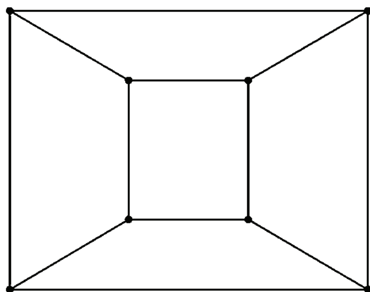


图　2

利用一些稍显复杂的方法可以证明，任意 9 个点都包含一个凸五边形，因此 $f(5)$ 等于 9。

2006 年，利用计算机，人们终于证明了 $f(6)=17$。目前，对于更大的 n，$f(n)$ 的值仍然都是未知的。人们猜测 $f(n)=2^{n-2}+1$，这个猜想是否正确，短时间内恐怕也无从得知了。

不管怎样，最后的结局真的很幸福。塞凯赖什和克莱因在结婚后的近 70 年里，先后到过上海和阿德莱德，最终在悉尼定居，期间从未分开过。2005 年 8 月 28 日，塞凯赖什和克莱因相继离开人世，相隔不到一个小时。



第三部分

几何的大厦

从灵机一动的想法开始，一点一点搭起整座大厦，最后竟能得出如此震撼的结论！



19. 尺规作图问题

有时候，比证明一个几何问题更有意思的，是怎样精确地把这个几何图形画出来。用尽可能简单的工具作出尽可能丰富的几何图形，无疑是一个非常吸引人的研究课题。事实上，从古希腊时代开始，人们就在研究几何作图，至今已经有 2000 多年的历史了。古希腊的数学家们敏锐地察觉到，直线和圆是最基本、最可信、亘古不变的几何概念，因而他们立下了一个规矩：几何作图只能使用直尺和圆规。这种选择很大程度上决定了今后平面几何研究的走向。平面几何理论的第一个框架是欧几里得的《几何原本》，其中有很多命题都是在讲如何作图的。《几何原本》的第一个命题就是一个作图问题，即如何作一个等边三角形。

以已知线段 AB 为边，作一个等边三角形（见图 1）。

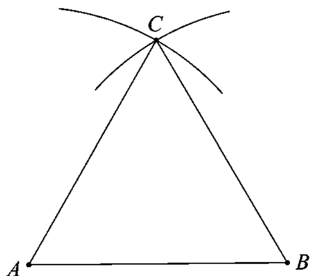


图　1

以 A 为圆心，AB 为半径作一条弧。显然，这条弧上的所有点到 A 的距离都等于 AB 的长。以 B 为圆心，AB 为半径再做一条弧。显然，这条弧上的所有点到 B 的距离都等于 AB 的长。那么，这两条弧的交点 C 就满足 $AC=BC=AB$，它就是等边三角形的第三个顶点。再用直尺把 AC、BC 连接起来，等边三角形就画好了。



关于尺规作图的概念，有两点值得大家注意。首先，直尺并不是普通的直尺——可能大家都知道——它是一条没有刻度的直尺。你不能用它测量已知线段的长度，更不能用刻度尺直接量出已知线段的中点在哪里。

另外，圆规也不是普通的圆规——这个估计就很少有人知道了——它是一个“松”了的圆规，只有作圆时才能固定张角，一旦离开纸面两脚便会“啪”的一声自动合拢。换句话说，在尺规作图问题中，圆规不能用来转移长度，只能作出以一点为圆心，以该点到另一点的距离为半径的圆。

不过，有没有这个限制关系并不大，因为《几何原本》的第二个命题就保证了，松圆规也能当作普通圆规用。

以已知点 A 为圆心，以已知线段 BC 的长度为半径作圆（见图 2）。

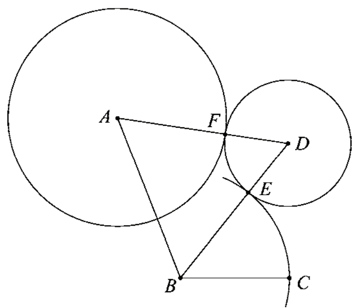


图 2

首先，利用前面的方法，作出等边三角形 ABD。然后，以 B 为圆心，BC 为半径作弧，交 BD 于 E；再以 D 为圆心，DE 为半径作弧，交 AD 于 F。由于 $AD=BD$，并且 $DF=DE$，因此 $AF=BE$；而 BE 又等于 BC，因此 AF 就等于 BC。现在，我们只需要以 A 为圆心，AF 为半径作圆就可以了。

每次需要用圆规转移线段长度时，我们都可以用上面这一招。这样一来，圆规就有用多了。

让我们先来看看，尺规作图是如何完成一些最基本的几何构造的。

作已知角的角平分线（见图 3）。

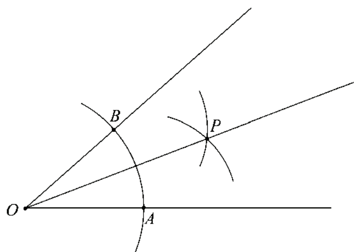


图　3

以 O 为圆心，任意长度为半径作圆，与角的两边分别交于 A、B 两点。再以任意长度为半径，分别以 A、B 两点为圆心作圆弧，两圆弧交于点 P。可以证明，$\triangle OBP$ 和 $\triangle OAP$ 是全等的，因而 OP 就是这个角的角平分线。

作垂直平分线的方法则更加简单。

作已知线段 AB 的垂直平分线（见图 4）。

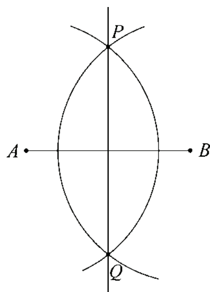


图　4

以适当长度为半径，分别以 A、B 两点为圆心作圆，两圆交于 P、Q 两点。由对称性，PQ 的连线就是线段 AB 的垂直平分线。注意，在作圆时，半径不能取得太短，否则两圆有可能没有交点。因此，我们用“适当长度”一词代替了“任意长度”。当然，更简便的做法是，直接取 AB 为半径长。

注意，这条垂直平分线与线段 AB 的交点，正是 AB 的中点。因此，我们也就有了尺规作图找出已知线段中点的办法。利用垂直平分线的作法，我们还可以完成下面这个操作。



过已知点 A 作已知直线的垂线（见图 5）。

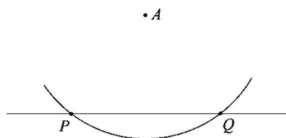


图　5

以 A 为圆心，以适当长度为半径作圆，与已知直线交于 P、Q 两点。然后，只需要套用刚才的方法，作出 PQ 的垂直平分线即可。在这里，取“适当长度”也是为了保证圆与直线有交点。

如果 A 恰好在直线上，上面的方法同样适用。

这又解决了下面这个问题：

过已知点 A 作已知直线的平行线（见图 6）。

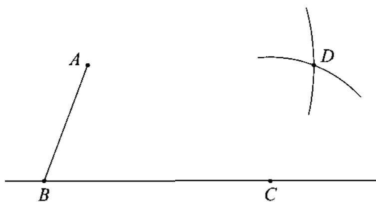


图　6

我们可以利用上面的方法，先过 A 点作已知直线的垂线，再过 A 点作这条垂线的垂线。

这里还有一个更巧妙的方法：在已知直线上任取两点 B、C，然后以 A 为圆心 BC 为半径作弧，再以 C 为圆心 AB 为半径作弧，把两条圆弧的交点记作 D。由于 $AD=BC$，$AB=CD$，因此四边形 $ABCD$ 是平行四边形，所以 AD 就是 BC 的平行线。

下面是一个看上去更加困难的问题。

过已知点 A 作已知圆 O 的切线（见图 7）。

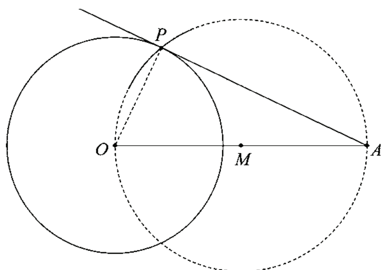


图　7

先作出 OA 的中点 M。然后，以 M 为圆心，以 OM 为半径作圆弧，与圆 O 交于点 P。AP 就是圆 O 的切线。这是因为，AO 是圆 M 的直径，因此 $\angle APO=90°$，这说明 AP 垂直于圆 O 的半径，也就说明了 AP 是圆 O 的切线。

当然，我们还有很多看上去更加困难的问题，其中最经典、最困难的作图问题可能要数阿波罗尼斯（Apollonius）问题了：作出一个圆，使得它与三个已知圆都相切（见图 8）。

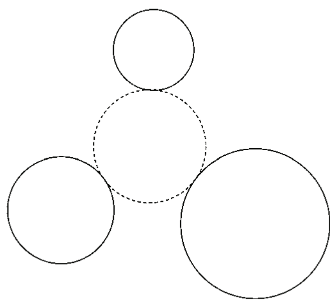


图　8

这个问题是古希腊数学家阿波罗尼斯提出并解决的，只不过他的解决方法已经失传了。16 世纪末，有数学家曾利用圆锥曲线的交点找到了阿波罗尼斯问题的解法，不过这并不算尺规作图的方法。1600 年，法国数学家弗朗索瓦 · 韦达（François Viète）给出了一个真正的尺规作图方案，才让阿波罗尼斯问题终于有了一个令人满意的答案。据推测，弗朗索瓦 · 韦达给出的方法很可能就是当年阿波罗尼斯自己的方法。到



现在，人们用代数、反演等不同的方法，已经找到了阿波罗尼斯问题的好几种更加漂亮的解法。看来，阿波罗尼斯问题也不是一个大问题了。

然而，一些看起来并不困难的几何构造，却怎么也没法用直尺和圆规做出来。古希腊人总结了三个真正的几何作图难题。

化圆为方：作出一个正方形，使得其面积和已知圆相等。

倍立方体：作出一个立方体，使得其体积是已知立方体的两倍。

三等分角：将任意一个已知角三等分。

注意，在平面几何中，我们显然是画不出立方体的。因此，在第二个问题中，我们已知的和要作的其实是立方体的棱长罢了。因此，第二个问题也就可以理解为，作出一条线段，使其长度是已知线段的 $\sqrt[3]{2}$ 倍。如果把已知线段看作单位长的线段的话，问题可以进一步简化为，作一条长为 $\sqrt[3]{2}$ 的线段。

1837 年，法国数学家皮埃尔·旺策尔（Pierre Wantzel）证明了一个惊人的结论：倍立方体永远不可能用尺规作图完成，换句话说只用直尺和圆规永远也不能作出长为 $\sqrt[3]{2}$ 的线段来。为了解释尺规作图为什么无法作出 $\sqrt[3]{2}$，我们先来看看尺规作图能够作出些什么。

首先，很容易想到，给定一条单位长的线段后，我们就能作出长度为任意正整数的线段来。

将已知线段 AB 延长到原来的 2 倍、3 倍、4 倍……（见图 9）。

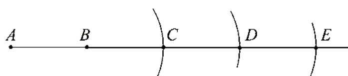


图 9

先用直尺延长 AB，然后以 B 为圆心 AB 为半径画弧，与直线交于点 C；再以 C 为圆心 AB 为半径画弧，与直线交于点 D……如此重复下去，我们便能得到长度为任意正整数的线段。

我们能把任意线段延长到原来的 n 倍，那有办法把它缩小到原来的 $\frac{1}{n}$ 吗？有！

将已知线段 AB 进行 n 等分（见图 10）。

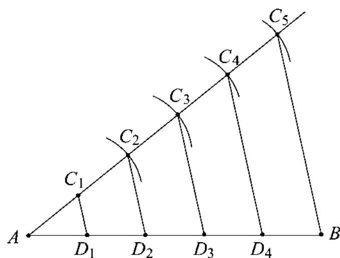


图 10

以 A 为端点，向任意方向作任意长度的线段 AC_1。然后，利用上面的方法，将 AC_1 延长到 2 倍、3 倍、4 倍……依次得到 $C_2, C_3, \cdots, C_n$。然后，把 C_n 与 B 相连，过 $C_1, C_2, \cdots, C_{n-1}$ 分别作 C_nB 的平行线，与 AB 分别交于 $D_1, D_2, \cdots, D_{n-1}$。显然，这些点就是线段 AB 的 n 等分点（图 10 演示的是 $n=5$ 的情形）。

这样一来，把单位线段先扩大到原来的 n 倍，再缩小到 $\frac{1}{m}$，就能得到长为 $\frac{n}{m}$ 的线段，从而便可以作出长度为任意有理数的线段了。

不但如此，我们还可以利用直尺和圆规，完成数与数之间的运算。下面就是把两个线段的长度加在一块儿的方法。

已知线段 AB 的长度为 a，线段 CD 的长度为 b，作一条长度为 $a+b$ 的线段（见图 11）。

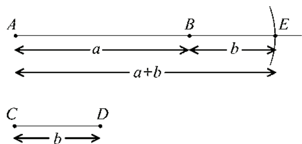


图 11

延长 AB，然后以 B 为圆心 CD 为半径作圆，与 AB 的延长线交于点 E（更常见的说法则是，在 AB 的延长线上截取 $BE=CD$）。AE 的长度就是 $a+b$。

类似地，我们也能用尺规作图实现两数相减（见图 12）。



已知线段 AB 的长度为 a，线段 CD 的长度为 b，作一条长度为 $a-b$ 的线段（假设 $a>b$）。

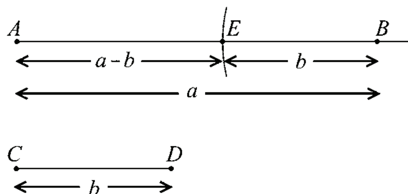


图 12

类似地，直接在线段 AB 内截取 $BE=CD$，则 AE 就等于 $a-b$。

下一个问题就比较有挑战性了：尺规作图能实现两数相乘吗？答案仍然是肯定的。不过，我们需要假设已经事先给定了一条单位长的线段。否则，我们无法确定出已知线段所表示的长度值，它们的乘积也就没有意义了。

已知线段 AB 的长度为 a，线段 CD 的长度为 b，作一条长度为 $a \cdot b$ 的线段（见图 13）。

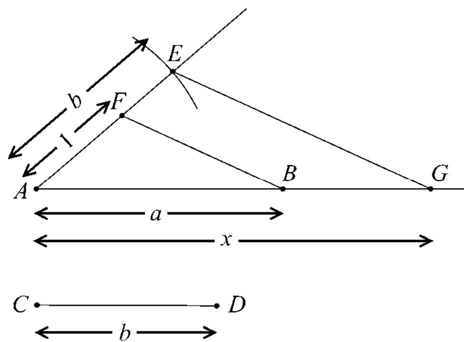


图 13

以 A 为端点，向任意方向作射线。用圆规在这条射线上截取 $AE=CD$，截取 $AF=1$。过 E 作 FB 的平行线，与 AB 的延长线交于点 G。如果把 AG 的长度记作 x，注意到图中大小两个三角形相似，则有 $a : x = 1 : b$，于是 $x = a \cdot b$。



图 13 所示的是 $b \geqslant 1$ 的情况。当 $b<1$ 时，这种方法仍然适用。

同理，我们也可以用尺规作图做除法。

已知线段 AB 的长度为 a，线段 CD 的长度为 b，作一条长度为 $\frac{a}{b}$ 的线段（见图 14）。

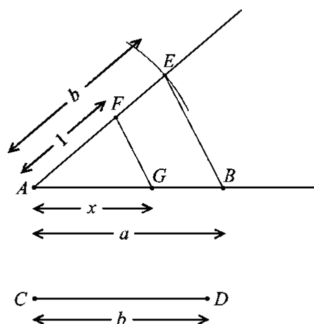


图　14

以 A 为端点，向任意方向作射线。用圆规在这条射线上截取 $AE=CD$，截取 $AF=1$。过 F 作 EB 的平行线，与 AB 交于点 G。如果把 AG 的长度记作 x，注意到图中大小两个三角形相似，则有 $x:a=1:b$，于是 $x=\frac{a}{b}$。

同样地，图 14 所示的是 $b \geqslant 1$ 的情况，事实上这种方法对 $b<1$ 的情况也是成立的。

现在，加减乘除四则运算都能用尺规作图完成了。那么，尺规作图能开平方吗？没有问题！

已知线段 AB 的长度为 a，作一条长度为 $\sqrt{a}$ 的线段（见图 15）。

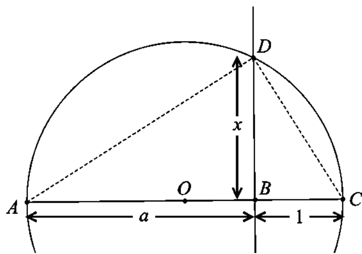


图　15



在 AB 的延长线上截取 $BC=1$。找出 AC 的中点 O。以 O 为圆心，OA 为半径作圆。过 B 作 AC 的垂线，与圆交于点 D。由于直径所对的圆周角是 90°，所以∠ADC=90°；再加上 BD 垂直于 AC，容易看出△ABD 与△DBC 相似。如果把 BD 的长度记作 x，则 $a:x=x:1$，也就是 $x=\sqrt{a}$。

到此为止，我们已经可以作出所有的有理数，可以对它们进行加减乘除和开平方运算，当然还可以对所得结果继续加减乘除，继续开平方。我们通常把从全体有理数出发，通过有限次加减乘除和开方运算可以得到的数都叫做“可构造数”（constructible number），意即它们可以用尺规作图构造出来。2、$\frac{2}{3}$、$\sqrt{2}$、$\sqrt{\frac{2}{3}}$、$\sqrt{2}+\sqrt{\frac{2}{3}}$、$\sqrt{2}+\sqrt{\frac{2}{3}}-1$、$\sqrt{\sqrt{2}+\sqrt{\frac{2}{3}}-1}$……这些数都是可构造数。

然而，尺规作图的极限也就到这里了。只用直尺和圆规，我们再也无法作出除了可构造数以外的其他数了。这是因为，在平面直角坐标系中，经过 (a_1,b_1) 和 (a_2,b_2) 两点的直线的方程是 $(b_1-b_2)x+(a_2-a_1)y+(a_1b_2-a_2b_1)=0$，以 (a_1,b_1) 为圆心且经过点 (a_2,b_2) 的圆的方程则是 $(x-a_1)^2+(y-b_1)^2=(a_2-a_1)^2+(b_2-b_1)^2$，这些方程的系数都是对已有点的坐标进行加减乘除运算得来的；而两个直线方程的公共解、两个圆的方程的公共解、一个直线方程和一个圆的方程的公共解，都能通过消元化简为一元一次方程或者一元二次方程，它们的解都可以用方程各系数之间的加减乘除和开方运算表示出来。一开始，我们只有一条单位长的线段，或者说只有 (0,0) 和 (1,0) 两个点。在此之后，不管怎么作图，产生的新交点的坐标始终都是已有点坐标通过四则运算和开方运算得到的，永远跳不出可构造数的圈子。既然每个点的坐标都是可构造数，而 (a_1,b_1)、(a_2,b_2) 两点间的距离公式是 $\sqrt{(a_1-b_1)^2+(a_2-b_2)^2}$，可见所有的距离值也都在可构造数的范围内。

尺规作图中偶尔会有“取任意点”、“取任意长度”的步骤（我们之前已经见过这样的例子了），此时我们也可以假定这些“任意点”的坐标以及“任意线段”的长度都是有理数。如果点或者长度真的是任意取的，这个假设也不会对作图过程带来任何影响。因而，在尺规作图中引入“任意”元素也不会让上面的证明失效。

但是，$\sqrt[3]{2}$ 不能表示成可构造数，这就证明了倍立方体是无法用尺规作图完成的。

皮埃尔·旺策尔所做的不仅仅是证明了倍立方体的不可能性，他还从尺规作图中最基本的元素出发，探索了一切可以用尺规作图构造的图形，建立了可构造数的理论，一举解决了数个尺规作图不可能性问题，倍立方体只是其中之一。

借助可构造数理论，三等分角的不可能性也很快得证。如图 16 所示，三等分已知角，本质上相当于已知 $\cos\theta$，求作 $\cos\left(\frac{\theta}{3}\right)$。利用三角函数我们可以推出，$\cos\theta = 4\cos^3\left(\frac{\theta}{3}\right) - 3\cos\left(\frac{\theta}{3}\right)$，因而三等分角相当于解这么一个三次方程，一般情况下是不能用尺规作图完成的。

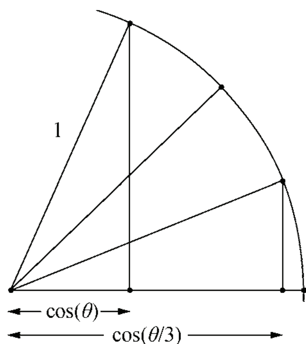


图　16

化圆为方本质上则相当于作出长为 $\sqrt{\pi}$ 的线段。1882 年，德国数学家林德曼（Ferdinand von Lindemann）证明了 π 是一个超越数（即它不是任何一个整系数多项式方程的解），给化圆为方也判了死刑。至此，古希腊三大几何作图难题的不可能性都被证明了。

不知道大家是否想过，如果当初古希腊人并没提出尺规作图，而是选用了别的作图工具，今天的几何学又会怎样呢？哪些原本简单的几何构造变得异常难以完成，哪些原本作不出的图形却可以轻而易举地画出来？换一块基石，重新搭建几何作图的大厦，无疑是一件激动人心的事情。



20. 单规作图的力量

其实，早在可构造数理论提出之前，就已经有人开始研究其他的作图工具了。1672年，丹麦数学家乔治・莫尔（Georg Mohr）找到了一种只用圆规就能完成尺规作图各项基本操作的方法，从而证明了这样一个惊人的事实：一切用尺规作图能够办到的事情，只用圆规也能办到。换句话说，尺规作图完全等价于单规作图。有趣的是，1797年，意大利数学家洛伦佐・马歇罗尼（Lorenzo Mascheroni）也独自发现了这个结论，因而这个结论被称为莫尔–马歇罗尼定理。

当然，有一件事是单规作图永远不可能办到的——只用圆规是没法画出直线的。不过，只要确定了直线上的两点，即使不画出这条直线来，这条直线的位置也已经确定了。倘若你能做到“眼中无线，心中有线”，单规作不了直线也不会带来什么障碍。而我们用直尺画直线的真正目的，其实是为了找出直线与其他线条的交点。如果不画出实际的直线，仅凭圆规就能找出这些交点的话，直尺也就没有用了。因此，为了证明单规作图可以完全代替尺规作图，我们只需要给出以下两个问题的单规作图方法。

(1) 已知 A、B 两点和圆 O，作出 A、B 所在直线与圆 O 的交点（如果有的话）（见图 1）。

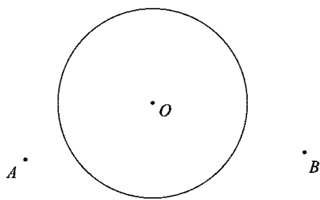


图 1

(2) 已知 A、B 两点和 C、D 两点，作出 A、B 所在直线和 C、D 所在直线的交点（见图 2）。

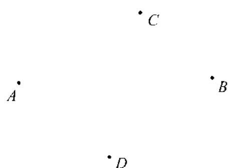


图 2

不过，要想“隔空”找出交点似乎并不是一件容易的事。下面，我们将从一些简单的问题出发，一步步探索单规作图的能力。先来看一个最基本的操作。

将线段 *AB* 延长到原来的 2 倍、3 倍、4 倍……（见图 3）。

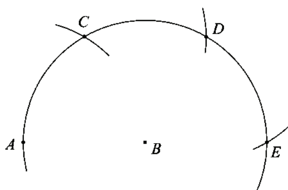


图 3

先以 *B* 为圆心，以 *AB* 为半径作圆。然后从 *A* 点出发，以 *AB* 为半径，在刚才所作的圆上连续截取 *AC*、*CD*、*DE*。由于△*ABC*、△*BCD*、△*BDE* 都是等边三角形，易知 *A*、*B*、*E* 三点在一条直线上，且 *AE* 的长度等于 *AB* 的两倍。重复此操作，便能将 *AB* 延长到任意整数倍的长度。

稍后，我们将在一个出人意料的地方用到上面这个操作。下面我们继续来看单规作图能够实现的另一项简单操作。

已知 *A*、*B* 两点以及点 *C*，作出点 *C* 关于 *A*、*B* 所在直线的对称点（见图 4）。

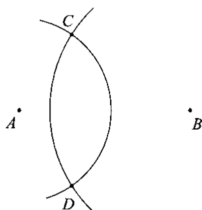


图 4



很简单，先以 A 为圆心，AC 为半径画弧，再以 B 为圆心，BC 为半径画弧，两弧的另一交点 D 即为所求。

学到了这一招后，我们立即有了求出直线与圆的交点的办法。

已知 A、B 两点以及圆 O，作出 A、B 所在直线和圆 O 的交点（见图 5）。

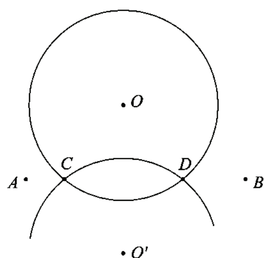


图 5

我们可以先作出 O 关于 AB 的对称点 O'，然后以 O' 为圆心，取圆 O 的半径为半径画弧，这条弧与圆 O 的交点 C、D 即为所求。如果这条弧与圆 O 不相交，那么 A、B 所在直线与圆 O 没有交点。

不过，上述方案有一个致命的缺陷：如果直线 AB 恰好经过圆心 O，这个方法就不能用了！看来，找出直线和圆的交点，依然没被完美解决。为了处理圆心正好在已知直线上的特殊情形，我们还得想想别的办法。

我们把这件事暂时放在一边，看看单规作图还能做些什么。

已知 A、B、C 三点，作出点 D，使得 $ABCD$ 是一个平行四边形（见图 6）。

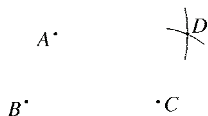
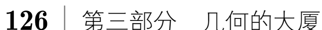


图 6

以 A 为圆心，BC 为半径作圆；再以 C 为圆心，AB 为半径作圆。两圆的交点 D 将同时满足 $AD=BC$ 以及 $AB=CD$，因而四边形 $ABCD$ 是平行四边形。



构造平行四边形给我们提供了很多有利的工具。我们可以借此完成下面这个重要的操作。

已知圆 O 和圆周上两点 A、B，作出弧 AB 的中点（见图 7）。

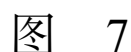


图 7

先作出平行四边形 $ABOC$ 和 $BAOD$。以 C 为圆心，CB 为半径作弧；以 D 为圆心，DA 为半径作弧，两弧交于点 E。最后，以 OE 作为半径，分别以 C、D 为圆心画圆弧，两弧交于 M、N 两点。下面我们证明，M、N 就是劣弧 AB 和优弧 AB 的中点。

由对称性，E、M、O 显然都在线段 AB 的垂直平分线上。如果 M 正好落在圆 O 上，它就一定是弧 AB 的中点。因此，如果把圆 O 的半径记作 r，为了说明 M 点是弧 AB 的中点，我们只需要说明 $OM=r$ 即可。

不妨把 AB 的长度记作 $2a$，把点 A 到 CD 的距离记作 h。由勾股定理可知，$a^2+h^2=r^2$。另外，由勾股定理还可以得到，$AD^2=(3a)^2+h^2=9a^2+h^2$。而 DE 是等于 AD 的，因此 DE^2 也等于 $9a^2+h^2$。于是，$OE^2=DE^2-OD^2=9a^2+h^2-4a^2=5a^2+h^2$。而 DM 是等于 OE 的，因此 DM^2 也等于 $5a^2+h^2$。因而 $OM^2=DM^2-OD^2=5a^2+h^2-4a^2=a^2+h^2=r^2$，说明 OM 就等于 r，这样便可得知 M 是劣弧 AB 的中点。

再由对称性可知，N 也就是优弧 AB 的中点。

有了单规平分弧的方法，我们也就能够求出特殊情况下直线和圆的交点了。



已知 A、B 两点以及圆 O，假设 A、B 所在直线正好经过圆心 O，作出 A、B 所在直线和圆 O 的交点（见图 8）。

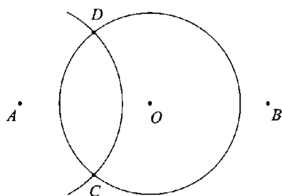


图 8

以 A 为圆心，适当长度为半径作圆，与已知圆 O 交于 C、D 两点。我们只需要利用刚才的办法，找出劣弧 CD 和优弧 CD 的中点即可。

这下，求出直线与圆的交点这一问题，就被我们彻底地解决了。我们的任务已经完成了一半。

为了求出直线与直线的交点，我们还需要做下面的准备工作。

已知长度分别为 a、b、c 的三条线段，求作一条长度为 d 的线段，使得 $a:b=c:d$。当然，单规作图是画不了线段的，已知的和求作的都仅仅是线段的两个端点（见图 9）。

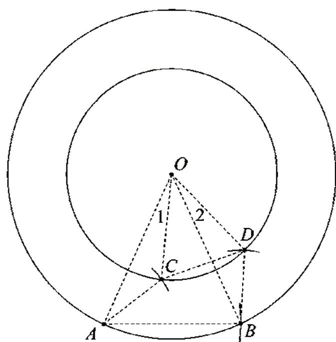


图 9

先考虑 $a>b$ 的情况。首先，以任意一点 O 为圆心，分别以 a、b 为半径作同心圆。在大圆上任取一点 A。以 A 为圆心，c 为半径画弧，与大圆交于点 B。因而，线



段 AB 的长度就是 c。现在，以适当长度为半径，分别以 A、B 为圆心作圆弧，与小圆分别交于 C、D 两点。因而，线段 AC 和线段 BD 的长度是相等的。我们下面将证明，CD 就是我们要求的线段。

注意到△ACO 和△BDO 的三条边都对应相等，因而它们全等，于是∠1=∠2。由此可知，∠AOB=∠COD。也就是说，△AOB 和△COD 是顶角相等的两个等腰三角形，它们显然是相似的。这就告诉我们，$AO:CO=AB:CD$，因此 CD 就是我们要求的线段。

如果 $a<b$，类似地，我们就在小圆上作 $AB=c$，在大圆上截得 $AC=BD$。根据同样的道理，CD 就是我们要求的线段。

不过，这里还有一个问题。在上述过程中，我们做了一个假设：在半径为 a 的圆中，我们能找到一条长度为 c 的弦 AB。但若 a 太小，c 太大的话，这样的弦可能就不存在了。稍作分析便可知道，只要 $c>2a$，我们永远也无法在圆上截出线段 AB 来。这该怎么办呢？此时，最早提到的“把线段延长整数倍”那一招就派上用场了。我们把线段 a 延长到一个足够大的倍数，比如 $n \cdot a$。然后用上面的方法作出满足 $(na):b=c:d$ 的线段 d。再把线段 d 也延长到 n 倍，就得到了我们本来要求的线段。

激动人心的时刻到了。我们即将完成莫尔–马歇罗尼定理的整个证明过程的最后一环。

已知 A、B 两点和 C、D 两点，作出 A、B 所在直线和 C、D 所在直线的交点（见图 10）。

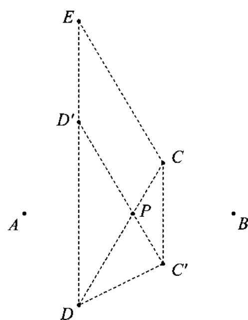


图 10



首先，作出 C 点关于直线 AB 的对称点 C'，以及 D 点关于直线 AB 的对称点 D'。我们要作的其实也就是 CD 和 $C'D'$ 的交点。不妨把这个点记作点 P。只有圆规，没有直尺，怎样作出点 P 呢？注意到，D 和 D' 到点 P 的距离是相等的，如果我们能作出一条长度等于 DP 的线段，便能以此长度为半径，分别以 D、D' 为圆心作圆弧，两圆弧的交点就是点 P 了。上哪儿找一条长度等于 DP 的线段呢？注意到 CC' 和 DD' 是互相平行的（因为它们都垂直于 AB），因而如果作出平行四边形 $CC'D'E$，那么 E 点将正好落在直线 DD' 上，并且 $\triangle DEC$ 与 $\triangle DD'P$ 相似。于是，$DE:DC=DD':DP$，其中 DE、DC、DD' 的长度都是已知的，我们就可以套用刚才的方法作出长度为 DP 的线段。以这个长度为半径，分别以 D、D' 为圆心作圆弧，两圆弧的交点便是所求的点 P 了。

这样一来，我们仅用一个圆规就能“模拟”尺规作图的所有基本操作，单规作图的能力也就与尺规作图一样了。



21. 锈规作图也疯狂

还能有比单规作图更疯狂的吗？有！

现在，让我们来考虑一种更坏的情况：假设我们不但没有直尺，而且连圆规也是坏的——这是一把生锈的圆规，两只脚已经被卡住了，只能画出单位半径的圆。在这样的条件下，哪些作图问题仍然能够被解决？

锈规作图相当困难，但并不是完全没有可能的。数学教授丹·佩多（Dan Pedoe）的一名学生惊奇地发现，给定两个点 A 和 B，如果它们的距离小于 2，我们可以非常简单地作出点 C，使得 $AC=BC=AB$（即$\triangle ABC$为等边三角形）。

如图 1，先以 A、B 为圆心分别作圆。由于它们之间的距离小于 2，因此两圆必然相交。以其中一个交点 P 为圆心作圆，分别交圆 A、圆 B 于点 M、N。最后，再分别以点 M、N 为圆心作圆，则圆 M 和圆 N 的交点即为所求点 C。由对称性，$\triangle CAB$ 一定是一个等腰三角形。另外，由对称性可知$\angle ACB=2\angle BCP$，而在圆 N 中，圆周角$\angle BCP$的度数又是圆心角$\angle BNP$的一半。因此，$\angle ACB$ 正好等于$\angle BNP$。由于$\triangle BNP$是等边三角形，我们可以立即得到$\angle ACB=\angle BNP=60°$，$\triangle ABC$ 也是一个等边三角形。

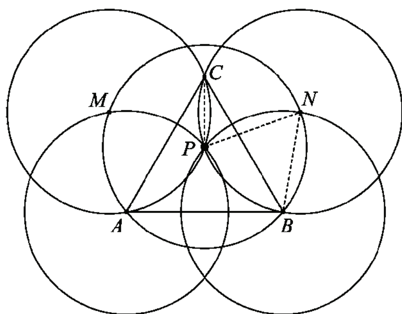


图　1



丹・佩多受到启发，在 1980 年的《数学难题》（*Crux Mathematicorum*）杂志上提出了这样一个问题：如果 A、B 是任意给定的两点，满足要求的点 C 仍然能够只用锈规作出吗？1982 年，佩多听闻了我国张景中和杨路给出的解答，并把它发表在了同一杂志上。让我们来看看，单用锈规是如何作等边三角形的吧。

首先，我们介绍锈规的第一个比较明显的用途：找出给定两点 A、B 间的一条由单位长线段首尾相接构成的折线段。方法很简单，如图 2，先作出圆 B，然后从点 A 出发，用锈规不断画弧并在弧上取点，然后以新的点为圆心继续画弧。总有一个时候，会有某个圆弧与圆 B 相交，此时我们所需要的折线段也就找到了。当然，由于我们没有直尺，我们是无法真正画出这条折线段的。不过这没什么，和单规作图一样，我们需要的只是这些点的位置。

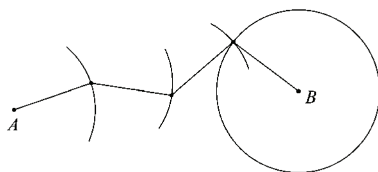


图 2

给定 A、B、C 三点，我们可以利用折线段构造法巧妙地作出平行四边形 $ABDC$。如图 3，首先作出从 A 到 B 的折线段，再作出从 A 到 C 的折线段，然后顺次作出一个个边长为 1 的菱形，最终得到的点 D 就是所求的点。注意到，菱形都是平行四边形，因此我们所作的实际上是把折线段 AB 一点点平移到了折线段 CD，自然也就相当于把线段 AB 平移到了线段 CD，因而四边形 $ABDC$ 显然是一个平行四边形。

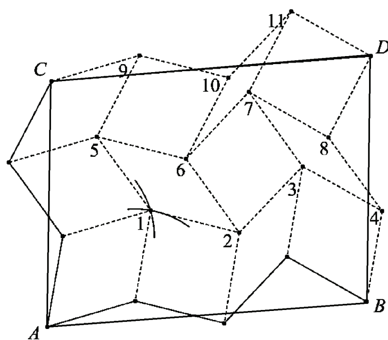


图 3

好了，我们已经慢慢地接近目标了。考虑这样一个作图问题：已知等边三角形 PAB 和等边三角形 PCD，能否只用锈规找出点 E，使得 BDE 也是一个等边三角形？如图 4，事实上，这个 E 点恰好就是使得四边形 $APCE$ 为平行四边形的那个点，借助上面的方法我们可以轻易作出 E 点的位置。利用初等平面几何知识不难证明，如果 $APCE$ 是平行四边形，则△ABE、△PBD、△CED 全等，从而△BDE 必然是一个等边三角形。详细的证明过程就留给大家自己去完成了。

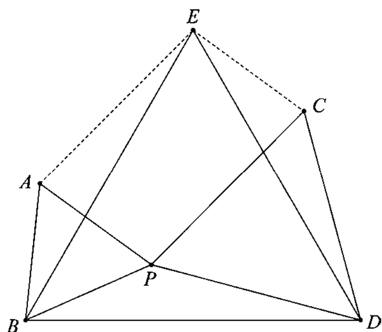


图　4

回到我们最初的问题，给定 A、B 两点后，我们可以像图 5 那样，先作出一条由单位长线段构成的折线 $A\text{-}P_1\text{-}P_2\text{-}\cdots\text{-}P_n\text{-}B$，进而作出 $n+1$ 个边长为 1 的等边三角形。然后，一次次套用作平行四边形的方法，作出 T_1, T_2, …等一系列的点，不断将两个小的等边三角形合成一个大的三角形。最后的 T_n 就是我们所求的 C 点，它使得△ABC 恰为一个等边三角形。至此，我们的问题已经圆满地解决了！

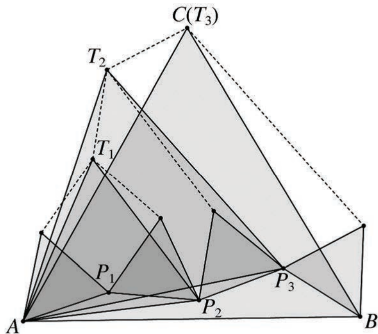


图　5



那么，给定 A、B 两点，能够作出正方形 $ABCD$ 吗？能够作出线段 AB 的中点 M 吗？侯晓荣等人对此做了进一步研究，利用复平面方法证明了一个非常强大的定理：事实上，从给定两点 A、B 出发，一切尺规作图能够完成的操作，锈规作图都能做到！1987 年，他们把这一成果撰写成《锈规作图论》一文，发表在了《中国科学技术大学学报》上。

值得注意的是，“从给定两点出发”这一条件必不可少。在有多个已知点的情况下，锈规作图的能力还有待研究。

22. 火柴棒搭成的几何世界

到目前为止，我们的讨论仍然没有跳出直尺和圆规的框架。这次，让我们彻底摆脱传统的束缚，来看一个全新的作图方法：火柴棒作图。

我曾经看到过这样一个智力题：如何用火柴棒准确地搭出一个正方形？注意，由于没有任何工具可以让两根火柴棒拼成一个 90°角，因此用四根火柴棒随意摆出一个四边形，最多也只能是个菱形。要想拼出一个正方形，我们还得想些奇招来。

一个经典的做法如图 1 所示。先摆出线段 *AB*，然后我们将会确定线段 *AK* 的位置，使得两条线段成 90°角。在 *AB* 上随意找一个点 *C*，以 *AC* 为底搭出两个腰为 1 的等腰三角形 *DAC* 和 *EAC*。容易看出，*D*、*E* 是关于 *AB* 对称的两个点。搭建一系列等边三角形△*ADF*、△*AFG*、△*AGH*，确定出 *D* 关于 *A* 点的对称点 *H*。这样，*H*、*E* 两点就关于 *AK* 轴对称了。再搭一个等边三角形 *AIE*，则 *I*、*G* 两点也关于 *AK* 对称。因此，*HG* 和 *IE* 的交点 *J* 就在 *AK* 上，自然 *AK* 的位置也就确定出来了。重复执行以上操作，我们便能完成以 *AB* 为边的整个正方形。

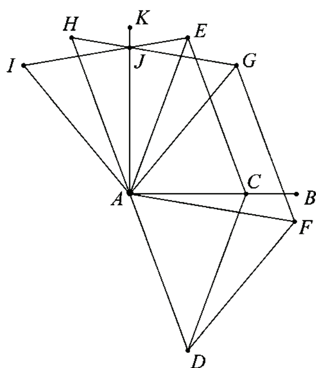


图　1



受此启发，我们自然而然地想到了这样一个问题：火柴棒的几何作图能力到底有多强？我们能仅凭借火柴棒找出线段的中点吗？我们能仅凭借火柴棒搭出一个正五边形吗？1939 年，道森（T. R. Dawson）在一篇论文中证明了一个惊人的结论：火柴棒作图与尺规作图的能力也完全一样！换句话说，用尺规作图能够确定的点，用火柴棒作图也能确定；而尺规作图办不到的事，火柴棒作图也没法办到。也就是说，和单规作图一样，火柴棒作图完全等价于尺规作图！

为了证明这一结论，我们首先得给火柴棒作图下一个定义。我们约定，用火柴棒作图时只允许以下四种基本操作，它们就是火柴棒几何中的“公理”。

(1) 给定一点 A，可以作一条通过 A 的单位长线段，或者以 A 为端点的单位长线段。

(2) 给定距离不超过单位长的两点 A、B，可以作一条通过 A、B 的单位长线段，或者以 A 为端点过 B 的单位长线段。

(3) 给定距离不超过单位长的两点 A、B，可以以 AB 为底作一个腰为单位长的等腰三角形 ABC。

(4) 给定点 A 和与其距离不超过单位长的直线 l，可以作一条以 A 为端点，另一端点在 l 上的单位长线段。

其中，公理 4 非常有用，我们将在后面反复用到它。

有了这些公理，我们便可以一步一步搭建火柴棒的几何世界了。先来看一个最基本的操作：延长一条线段。

延长一条线段。

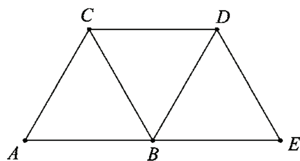


图 2

如图 2 所示，搭出一系列等边三角形，我们便能把 AB 延长到 AE。重复这样的操作，便可将一条线段无限延长。

注意到线段 CD 与 AB 平行且相距 $\frac{\sqrt{3}}{2}$ 个单位，因此我们还得到了一个非常有用的工具：将给定线段平移 $\frac{\sqrt{3}}{2}$ 个单位。

下面我们再来看一个基本操作。

找出长度小于单位长的线段的中点。

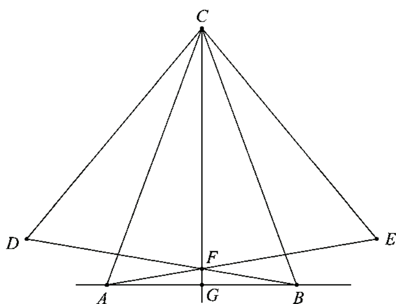


图　3

如图 3 所示，AB 为已知线段。先作等腰三角形 ABC，再作等边三角形 BDC 和 AEC。BD 和 AE 的交点 F 就在等腰三角形的中线上。CF 的延长线与 AB 的交点就是我们所求的点 G。

不过，当线段 AB 的长度等于或大于单位长时，这种方法就不能用了。

注意到，由于 CG 还平分了 $\angle ACB$ 和 $\angle DCE$，因此我们相当于有了一个平分不超过 120°且不等于 60°的角的办法。另外，由于 CG 还是 AB 的垂线，因此我们又有了过点 C 向已知线段作垂线的方法——先利用公理 4 摆出线段 CA 和 CB，再找出 AB 的中点。即使 C 点离已知线段很远，垂线照样能作出，因为我们可以将已知线段不断平移 $\frac{\sqrt{3}}{2}$ 个单位，让它与 C 的距离足够近。不过，这里还是有一种特殊的情况：若 C 与已知线段的距离恰好是 $\frac{\sqrt{3}}{2}$ 的整倍数，这么做就不行了。

当线段 AB 的长度等于单位长时，我们可以用下面的办法找出中点。

找出长度等于单位长的线段的中点。



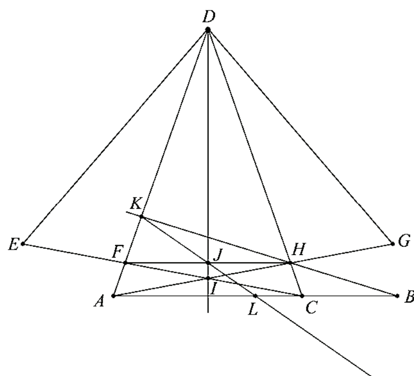


图 4

如图 4 所示，假如 AB 是一条长度恰为单位长的已知线段。首先在 AB 上任取一点 C，然后作等腰三角形 ADC。作等边三角形 CED，与 AD 交于 F；作等边三角形 AGD，与 CD 交于 H；CE 和 AG 交于点 I。那么，DI 与 FH 的交点 J 就是 FH 的中点。BH 与 AD 交于点 K，KJ 与 AB 交于点 L，于是我们就成功地把 FH 的中点转移到了 AB 的中点。

这个构造弥补了我们之前留下的空缺。现在，我们不但能平分恰为 60°的角，也能引出长度恰为 $\frac{\sqrt{3}}{2}$ 的整倍数的垂线了。

利用这些基本操作，我们可以实现一些更复杂的几何构造了。

过已知线段外的一点，作已知线段的平行线。

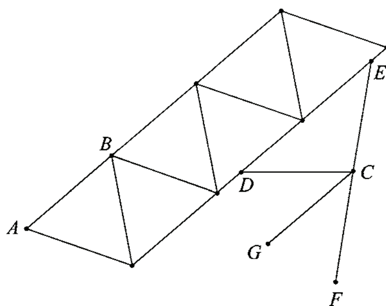


图 5

如图 5 所示，不断平移已知线段 AB，直到它离点 C 足够近。以 C 为端点，利用公理 4 引单位长线段 CD、CE。反向延长 CE 到 F，容易证明$\angle DCF$ 的平分线 CG 就与 AB 平行。

注意，虽然我们现在只能平分小于 120°的角，但好在，只要把线段 AB 平移到了离 C 点距离小于 $\frac{\sqrt{3}}{2}$ 的地方后，$\angle DCF$ 总是小于 120°的。

这就解决了下面这个大难题。

找出距离大于单位长的两点的中点。

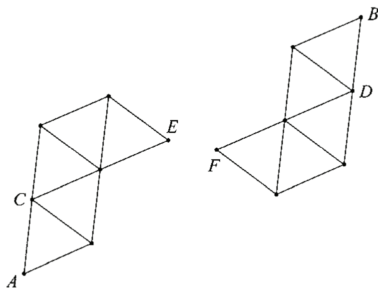


图　6

如图 6，已知很远的两点 A、B。向任意方向作单位长线段 AC，过 B 作它的平行线段 BD。利用一系列等边三角形，构造逐渐向中间靠拢的中心对称图形，直到出现距离不超过单位长的对称点 E、F。EF 的中点也就是 AB 的中点。

既然我们能找到任意线段的中点，平分大于 120°的角也就不成问题了。

好了，准备工作基本结束，下面我们就来说明火柴棒作图与尺规作图的等价性。注意到，火柴棒作图的四项基本操作都能用尺规作图实现，因此火柴棒作图是尺规作图的子集。为了说明尺规作图同时也是火柴棒作图的子集，我们只需要用火柴棒实现这样三个基本操作：作出过两点的直线、作出直线和圆的交点，作出圆和圆的交点。这样，火柴棒便能完全代替直尺和圆规了。

我们先来看最简单的一个：作出过两点的直线。

作出过 A、B 两点的直线。



为了连接 AB，首先找出 AB 的中点 C，然后找出 AC 的中点 D，BC 的中点 E……如此下去，直到 AB 之间有足够多的点，相邻点的距离都小于单位长度。这样，我们便可以用火柴棒连接很远的两点了。

作出直线和圆的交点就比较复杂了。

作出直线和圆的交点。

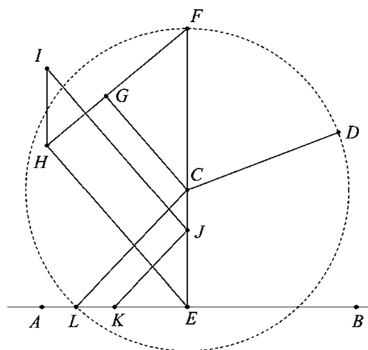


图　7

如图 7 所示，给定点 A、点 B、圆心 C 以及圆周上一点 D，我们需要找到直线 AB 与（隐形的）圆 C 的交点 L。过 C 作 $CE \perp AB$。在 CE 的反向延长线上截取 $CF=CD$（这是可以办到的，比如先作 $\angle DCF$ 的角平分线，再过 D 作角平分线的垂线；后面还会反复用到这个技巧）。向任意方向作单位长度线段 FG。过 E 作 CG 的平行线，交 FG 延长线于 H。过 H 作 EC 的平行线，截取 $HI=HG$。作 $IJ /\!/ HE$。

此时，图中的一系列平行线和等长线段告诉我们，$CE:CD=CE:CF=HG:GF=HI:GF=JE:GF$，而 CE 是小于半径 CD 的，因此 JE 是小于单位长线段 GF 的。于是，我们便可以利用公理 4 作单位长线段 JK。最后，过 C 作 JK 的平行线，把它与 AB 的交点记作点 L。由于 $CE:CD=JE:GF=JE:JK=CE:CL$，可见 CL 正好等于圆的半径长 CD，因此 L 点就是我们要求的圆与直线的交点。

最后，我们只剩下一步了：用火柴棍作出圆与圆的交点。

作出圆和圆的交点。

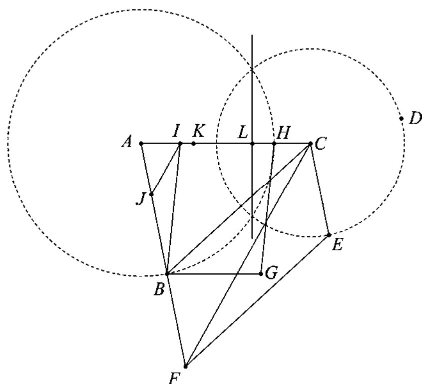


图 8

如图 8 所示，已知圆心 A 和圆周上一点 B，圆心 C 和圆周上一点 D，我们想要找出这两个圆的交点。由于我们已经能作直线与圆的交点了，因此为了作出两圆的交点，只要能找出公共弦所在直线即可。而公共弦与连心线垂直，因此我们只需要找出公共弦与连心线的交点 L 即可。不妨把圆 A 的半径记作 a，把圆 C 的半径记作 c，再在连心线上找出 $LK=LC$，则由勾股定理可得 $a^2-AL^2=c^2-CL^2$，即 $a^2-c^2=AL^2-CL^2$，再利用平方差公式可得 $(a+c)(a-c)=AC\cdot AK$。也就是说，AK 就等于 $\dfrac{(a+c)(a-c)}{AC}$。我们将利用这个关系找出 K 点来。

过 C 作 AB 的平行线，截取 $CE=CD$。作 $EF /\!/ CB$，则 AF 就等于 $a+c$。过 B 作 AC 的平行线，截取 $BG=BF$。截取 $AH=AB$，然后作 $BI /\!/ GH$，AI 就等于 $a-c$。作 $IJ /\!/ CF$，由 $\triangle AIJ$ 与 $\triangle ACF$ 相似可知 $AJ:AI=AF:AC$，因此 $AJ=\dfrac{AF\cdot AI}{AC}=\dfrac{(a+c)(a-c)}{AC}$。最后，只需要截取 $AK=AJ$，再找出 CK 的中点 L，问题就圆满解决了。

这样一来，所有尺规作图能够办到的事情，只用火柴棒也能办到，一切火柴棒作图问题都被终结掉了。不过，对火柴棒几何的研究还远未结束。如何简化作图过程，作出指定图形最少需要多少根火柴棒……这些悬而未决的问题都还有待人们继续探索。



23. 折纸的学问

我们研究了几个很容易想到的另类作图工具。但到目前为止，我们还没有发现哪种几何作图模型的作图能力可以超越尺规作图。难道，尺规作图真的就是最强大的作图工具了吗？当然不是。这可能有些令人难以置信，一个看上去比尺规作图更“低端”的作图方法，其能力竟然远远超过了尺规作图。这种方法就是——折纸。

1980 年，北海道大学的阿部恒（Hisashi Abe）发现，用折纸法可以三等分任意角，而这是尺规作图无法办到的。

假设我们要三等分的角是 $\angle XAY$。如图 1，把 $\angle XAY$ 放在矩形纸张的一个直角上，AY 靠着纸的边缘，AX 落在纸张内部。在纸张的另一直角边上确定两点 P 和 Q 使得 $AP=PQ$。过 P 点作 AY 的平行线。现在，把纸折起来，让 Q 点恰好落在 AX 上，同时 A 点也恰好落在那条平行线上。不妨把 A、P、Q 的落点分别命名为 A'、P'、Q'，那么 AP' 和 AA' 就是 $\angle XAY$ 的三等分线。

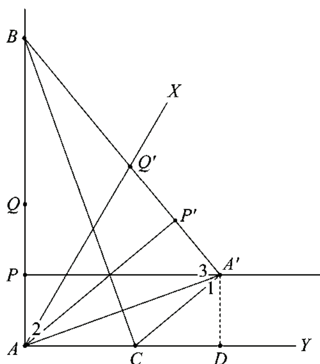


图 1

下面我们就来证明这一点。把折痕的端点分别记作 B 和 C，再过 A' 作 $A'D$ 垂直于

AY。由于 $A'D$ 平行于 PA，因此∠1=∠2；另外，$AB=A'B$，△BAA'是等腰三角形，于是∠2 又等于∠3。因此，∠1 就和∠3 相等。再加上 $A'D=PA=P'A'$，AA'是公共边，足以说明△$AA'P'$全等于△$AA'D$。这样，∠$AP'A'$ 就也是直角，从而∠$AP'Q'$也是直角。又注意到 $AP=PQ$ 即表明 $A'P'=P'Q'$，公共边 $AP'=AP'$，于是△$AP'A'$全等于△$AP'Q'$。以 A 为顶点的三个直角三角形都全等，因此对应的三个角相等。

折纸为什么会比尺规作图更强呢？要解答这个问题，首先我们得解决一个更基本的问题：什么叫折纸，折纸的游戏规则是什么？换句话说，折纸允许哪些基本的操作？大家或许会想到一些折纸几何必须遵守的规则：所有直线都由折痕或者纸张边缘确定，所有点都由直线的交点确定，折痕一律是将纸张折叠压平再展开后得到的，每次折叠都要求对齐某些已有几何元素（不能凭感觉乱折），等等。不过，这些定义都太"空"了，我们需要更加形式化的折纸规则。1991 年，藤田文章（Humiaki Huzita）指出了折纸过程中的 6 种基本操作（也可以叫做折纸几何的公理，示于图 2）。

(1) 已知 A、B 两点，可以折出一条经过 A、B 的折痕。

(2) 已知 A、B 两点，可以把点 A 折到点 B 上去（这并不难办到，不妨想象这张纸是透明的，所有几何对象正反两面都能看见，下同）。

(3) 已知 a、b 两条直线，可以把直线 a 折到直线 b 上去。

(4) 已知点 A 和直线 a，可以沿着一条过 A 点的折痕，把 a 折到自身上。

(5) 已知 A、B 两点和直线 a，可以沿着一条过 B 点的折痕，把 A 折到 a 上。

(6) 已知 A、B 两点和 a、b 两直线，可以把 A、B 分别折到 a、b 上。

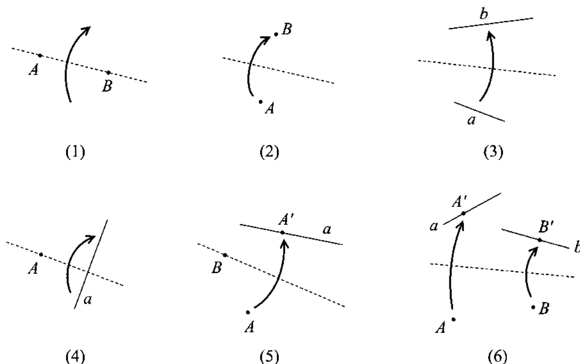


图　2



容易看出，它们实际上对应着不同的几何作图操作。例如，操作 1 实际上相当于连接已知两点，操作 2 实际上相当于作出已知两点的连线的垂直平分线，操作 3 则相当于作出已知直线的夹角的角平分线，操作 4 则相当于过已知点作已知直线的垂线。真正强大的则是后面两项操作，它们确定出来的折痕要满足一系列复杂的特征，不是尺规作图一两下能作出来的（有时甚至是作不出来的）。正是这两个操作，让折纸几何有别于尺规作图，折纸这门学问从此处开始变得有趣起来。

更有趣的是，操作 5 的解很可能不止一个。如图 3，在大多数情况下，过一个点有两条能把点 A 折到直线 a 上的折痕。

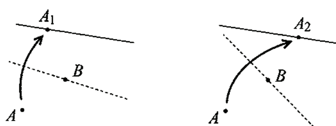


图 3

操作 6 则更猛，如图 4，把已知两点分别折到对应的已知两直线上，最多可以有三个解！

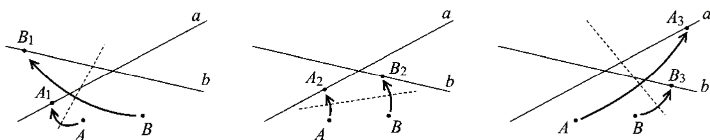


图 4

一组限定条件能同时产生三个解，这让操作 6 变得无比灵活，无比强大。利用一些并不太复杂的解析几何分析，我们能得出操作 6 有三种解的根本原因：满足要求的折痕是一个不可约的三次方程的解。也就是说，给出两个已知点和两条对应的已知直线后，寻找符合要求的折痕的过程，本质上是在解一个三次方程！

让我们来回顾一下尺规作图里的五个基本操作：

(a) 过已知两点作直线；

(b) 给定圆心和圆周上一点作圆；

(c) 寻找直线与直线的交点；

(d) 寻找圆与直线的交点；

(e) 寻找圆与圆的交点。

这五项操作看上去变化多端，但前三项操作都是唯一解，后两项操作最多也只能产生两个解。从这个角度来看，尺规作图最多只能解决二次问题，加减乘除和不断开方就已经是尺规作图的极限了。能解决三次问题的折纸规则，势必比尺规作图更加强大。

正因为如此，一些尺规作图无法完成的任务，在折纸几何中却能办到。这就回到了本节开头提到的问题：用折纸法可以实现三等分角，而这是无法用尺规作图办到的。

我们有更简单的例子来说明，用折纸法能完成尺规作图办不到的事情。前面我们讲过，“倍立方体”问题是古希腊三大尺规作图难题之一，它要求把立方体的体积扩大到原来的两倍，本质上是求作 2 的立方根。由于尺规作图最多只能开平方，因而它无法完成“倍立方体”的任务。但是，折纸公理 6 相当于解三次方程，解决“倍立方体”难题似乎游刃有余。

有意思的是，用纸片折出 2 的立方根比想象中的更加简单。取一张正方形纸片，将它横着划分成三等份（方法有很多，大家不妨自己想想）。然后，将右边界中下面那个三等分点折到正方形内部的上面那条三等分线上，同时将纸片的右下角顶点折到正方形的左边界。那么，纸片的左边界就被分成了 $\sqrt[3]{2}:1$ 两段。

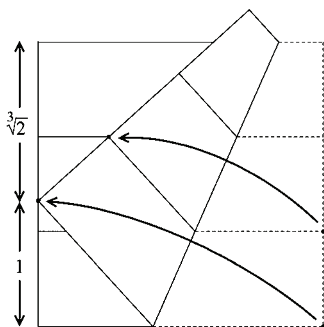


图　5

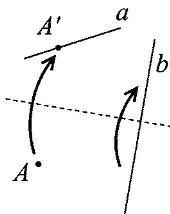
利用勾股定理和相似三角形建立各线段长度的关系，我们不难证明它的正确性。强烈建议大家自己动手算一算，来看看三次方程是如何产生的。



本文写到这里，大家或许以为故事就结束了，但 10 年以后（也就是 2001 年），事情又有了转折：羽鸟公士郎（Koshiro Hatori）发现，藤田文章的 6 个折纸公理并不是完整的。羽鸟公士郎给出了折纸的第 7 种操作。从形式上看，第 7 公理与已有的公理如出一辙，并不出人意料，很难想象这个公理整整十年里竟然一直没被发现。继续阅读之前，大家不妨先自己想想，这个缺失的操作是什么。这段历史背景无疑让它成为了一个非常有趣的思考题。

羽鸟公士郎补充的公理是：

(7) 已知点 A 和 a、b 两直线，可以把 A 折到 a 上，同时把 b 折到自身上。



(7)

图 6

后来，这 7 条公理就合称为藤田–羽鸟公理。在 2003 年的一篇文章中，罗伯特·兰（Robert J. Lang）对这些公理进行了一番整理和分析，证明了这 7 条公理已经包含折纸几何中的全部操作了。

罗伯特·兰注意到了，上述 7 项基本操作其实是由一些更基本的操作要素组合而成的，例如“把已知点折到已知线上”、“折痕经过已知点”等。说得更贴切一些，这些更加基本的操作要素其实是对折痕的“限制条件”。在平面直角坐标系中，折痕完全由斜率和截距确定，它等价于一个包含两个变量的方程。不同的折叠要素对折痕的限制力是不同的，例如“把已知点折到已知点上”就同时要求 $x_1' = x_2$ 并且 $y_1' = y_2$，据此可以建立出两个等量关系，一下子就把折痕的两个变量都限制住了。而“折痕经过已知点”则只包含一个等量关系，只能确定一个变量（形式上通常表示为与另一个变量的关系），把折痕的活动范围限制在一个维度里。

不难总结出，基本的折叠限制要素共有 5 个：

(1) 把已知点折到已知点上，确定 2 个变量；

(2) 把已知点折到已知线上，确定 1 个变量；

(3) 把已知线折到已知线上，确定 2 个变量；

(4) 把已知线折到自身上，确定 1 个变量；

(5) 折痕经过已知点，确定 1 个变量；

而折痕本身有 2 个待确定的变量，因此符合要求的折纸操作只有这么几种：(1)，(2)+(2)，(3)，(4)+(4)，(5)+(5)，(2)+(4)，(2)+(5)，(4)+(5)。但是，这里面有一种组合需要排除掉：(4)+(4)。在绝大多数情况下，(4)+(4)实际上都是不可能实现的。如果给出的两条直线不平行，我们就无法折叠纸张使得它们都与自身重合，因为没有同时垂直于它们的直线。

另外 7 种则正好对应了前面 7 个公理，既无重合，又无遗漏。折纸几何至此便有了一套完整的公理。

不过，折纸的学问远远没有到此结束。如果允许单次操作同时包含多处折叠，折纸公理将会更复杂，更强大。折纸的极限究竟在哪里，这无疑是一个非常让人振奋的研究课题。



24. 万能的连杆系统

在机器时代，作为机械构造的理论工具，连杆系统曾一度成为数学界中最热门的话题。所谓连杆系统，就是一些刚性的小杆在端点处以转轴的方式相连，形成的一个机械装置。固定某些顶点的位置之后，其余的动点就能画出一些有趣的轨迹。例如图1中的左图，固定杆 AB 的其中一个端点 A，则端点 B 将描绘出一个绕 A 点的圆周。

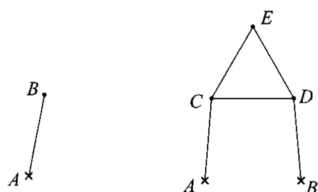


图 1

连杆系统最激动人心的，莫过于一些简单的连杆装置能够描绘出非常复杂的曲线。例如，图1的右图就是由五根相同长度的杆构成的连杆系统。固定 A、B 两个端点后，显然 C 和 D 描绘出的都是圆弧，但 E 点的轨迹就难以想象了。事实上，E 点的轨迹相当诡异，需要用一些复杂的代数语言才能描述（见图2）。

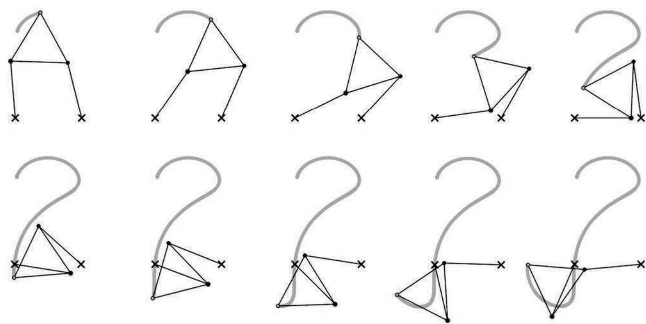


图 2



在连杆系统领域中，有一个困扰人类近百年的难题——利用连杆系统是否能画出直线来？当时看来，这个问题是如此地困难，人们甚至试图去证明，能画出直线的连杆系统压根儿是不存在的。1864 年，一位法国海军军官查尔斯–尼古拉斯 · 波赛利（Charles-Nicolas Peaucellier）发明了第一个能画出直线的连杆系统，在当时引起了极大的轰动。波赛利连杆系统的原理并不难理解，利用初中几何知识足以证明其正确性。

波赛利连杆是由 7 根杆组成的，如图 3，其中 $AC = AD = a$ ， $BC = CE = ED = DB = b$ ，OB 为任意长。固定 A 点和 O 点的位置，使得 OA 的距离恰好等于 OB，则 E 点将会描绘出一条垂直于 AO 的直线来。

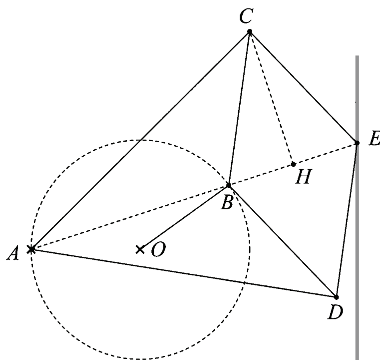


图　3

容易看出，A、B、E 三点在同一条直线上。我们首先说明，$AB·AE$ 是一个常数。过点 C 作 $CH \perp AE$，垂足为 H。于是

$$
\begin{aligned}
AB \cdot AE &= (AH + HE) \cdot (AH - HB) \\
&= AH^2 - BH^2 \\
&= (AC^2 - CH^2) - (BC^2 - CH^2) \\
&= a^2 - b^2
\end{aligned}
$$

结果是一个常数。

为什么 $AB·AE$ 为常数，就能保证 E 点的轨迹是一条直线呢？如图 4，过 A 点作出圆 O 的直径 AM，在射线 AM 上找出一点 N 使得 $AM·AN$ 也等于这个常数。由于 $AM·AN=AB·AE$，或者说 $\dfrac{AM}{AB} = \dfrac{AE}{AN}$，我们立即可知 $\triangle ABM$ 相似于 $\triangle ANE$，因此 $\angle ANE=\angle ABM=90°$，也就是说 EN 与 AN 始终垂直。这就证明了，E 点的轨迹确实是



一条与 AO 垂直的直线。

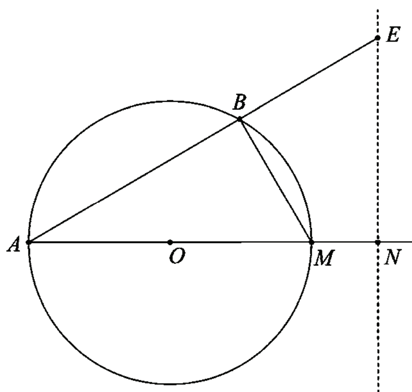


图 4

解决了连杆画直线的问题后，数学家们显然还不满足。很多迹象都表明，连杆系统比我们想象中的更强大，画出一些更奇怪的图形似乎不成问题。

有一个非常简单的构造几乎是瞬间增强了连杆系统的功能，让人们更加相信构造复杂连杆系统的可能性。虽然连杆系统要求杆与杆必须在端点处连接，但我们可以利用三角形的稳定性，把某根杆的一端直接接到另一根杆的中间。如图 5，虽然 AB 和 BC 是两根各自能绕着 B 转的杆，但简单地用三角形固定一下，AB 和 BC 将会变成一根杆 AC。利用这一基本构造，我们就能把杆的端点直接连在另一根杆的中间了。

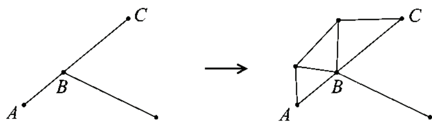


图 5

这一基本构造极大地激励了我们——我们何不像研究尺规作图一样，借助最基本的构造，构造出更实用的基本构造，逐渐搭建起连杆作图的大厦呢？1877 年，英国数学家艾尔弗雷德·肯普（Alfred Kempe）顺着这个思路研究下去，最后得出了一个惊人的结论：连杆系统不仅能画出直线和圆，还能画出双曲线、抛物线、椭圆，甚至半

立方抛物线、双纽线等复杂的曲线。事实上，任何代数曲线 $f(x,y)=\sum_i\sum_j C_{i,j}\cdot x^i\cdot y^j=0$ 都是可以用连杆系统画出的！

这个证明的基本思路是这样的。如图 6，首先，以 O 为端点构造两个菱形。利用两个波赛利连杆系统，我们可以让 x 点和 y 点始终沿着两条垂直的直线运动。固定 O 点后，我们就建立起了一个平面直角坐标系。接下来，我们需要把 y 点绕着原点顺时针旋转 90 度。假设菱形 $OCyD$ 的边长为 l，则构造连杆 $OC'=C'y'=y'D'=D'O=l$，$CC'=DD'=\sqrt{2}\cdot l$，这样我们就把 Oy 的长度转移到了 x 轴上。接下来，我们将用一系列连杆构造出一个点 T，使得 T 始终在坐标系中的 $(f(x,y),0)$ 的位置上。然后我们将构造出一个点 S，使得 S 始终在坐标系中的 (x,y) 位置上。最后，我们把 T 点的位置固定在 $(0,0)$，则 S 点就将描绘出 $f(x,y)=0$ 的图像来。

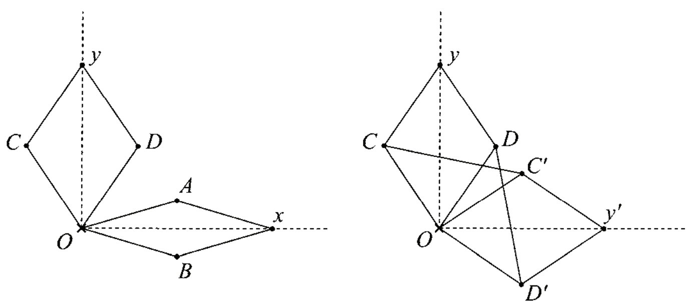


图　6

为了得到 $T(f(x,y),0)$，我们只需要实现对 x 轴上的点的以下四种操作：

(1) 把某个点的坐标加上一个常数 c；

(2) 把某个点的坐标乘上一个常数 c；

(3) 把两个点的坐标相加；

(4) 把两个点的坐标相乘。

前两个操作并不困难。如图 7，对于 x 轴上的某个点 p，为了得到点 $z=p+c$，只需要固定两个距离为 c 的点 A、B，并构造一系列平行四边形即可。为了得到点 $z=c\cdot p$，我们只需要构造一组相似三角形 OAp 和 OBz，使得 $OB=c\cdot OA$，$Bz=c\cdot Ap$。添加一个杆 pC 使得四边形 $ABCp$ 为平行四边形，以保证这两个三角形是相似的。注



意，在乘法器的构造中，我们用到了前面所说的基本构造，即杆的中间直接连接另一根杆。

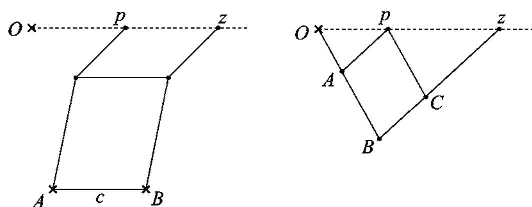


图 7

把两个变量相加也比想象中的容易。事实上，我们不但能在 x 轴上对两个动点做加法，还能直接实现一个更强的基本操作——对平面上的两个向量进行相加。如图 8，只需要构造一系列的平行四边形，容易看出四边形 $Opzq$ 也是一个平行四边形，向量 Oz 即是向量 Op 和 Oq 之和。

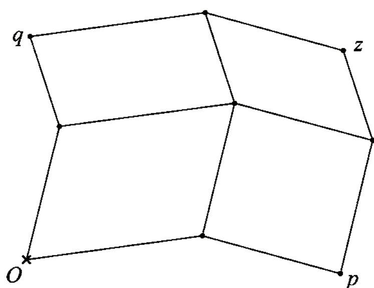


图 8

但是，对 x 轴上的两个变量进行相乘就有些麻烦了。注意到，由于 $p\cdot q=\dfrac{(p+q)^2-(p-q)^2}{4}$，因此只要能实现平方操作，我们也就有了实现乘法的方法。而由于 $\dfrac{1}{p-1}-\dfrac{1}{p+1}=\dfrac{2}{p^2-1}$，因此只要能实现倒数操作，我们也就有了实现平方的方法。在证明波赛利连杆系统的正确性时，我们已经证明了，在图 9 的连杆系统中有 $z\cdot p=a^2-b^2$，利用它我们便能实现 $z=\dfrac{a^2-b^2}{p}$。取 a 、b 为适当的值，我们就能得到 p 的倒数了。

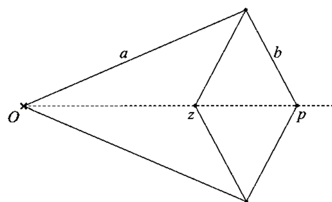


图　9

由于 x 和 y' 都已经在 x 轴上了，利用上面的这些基本操作，我们便能得到 $T(f(x,y),0)$。另外，利用向量加法器，我们可以得到 Ox 和 Oy 的向量和 $S=(x,y)$。将 T 点的位置固定在原点 O 处，S 的轨迹就是 $f(x,y)=0$ 的图像了。

肯普的结论最令人惊讶的地方莫过于，由于各种曲线都能用代数曲线近似地描述，因此连杆系统几乎是万能的了。因此，如果足够有耐心，你甚至能构造一个连杆系统，让它签出你的大名来！



25. 探索图形剪拼

大家或许都见过这种类型的谜题：

图 1 是由 5 个相同的小正方形组成的图形，请你把它裁剪成三块，然后拼成一个大正方形。

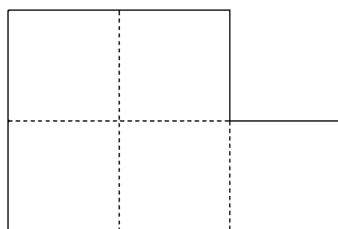


图 1

盲目尝试并不是一个好方法。注意到，把图 1 剪拼成一个正方形后，面积仍然是 5，因而它的边长一定是 $\sqrt{5}$ ，也就是说我们需要构造长度为 $\sqrt{5}$ 的线段。想到这一点，答案很快就出来了（见图 2）。

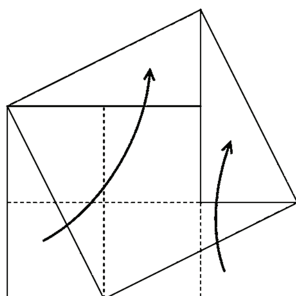


图 2

受此启发，我们可以利用图 3 的办法，把任意两个正方形剪成五块，然后拼成一个大正方形。

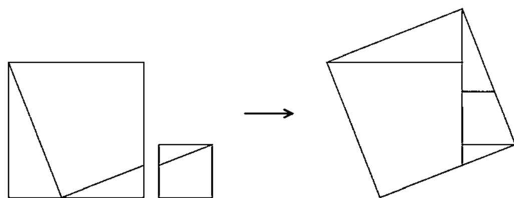


图 3

这给我们带来了图形剪拼谜题中一个一般的结论。

结论 1 两个正方形可以剪拼成一个大正方形。

有趣的问题来了：如果给你三个正方形，你还能把它们剪拼成一个大正方形吗？答案是肯定的。借助“二方合一”法，我们可以立即得到把三个正方形剪拼成一个正方形的方法：只需要先用这个方法把其中两个正方形剪拼成一个正方形，然后又一次使用“二方合一”把它与剩下的那个正方形再合成一个大正方形。显然，不管一开始正方形有多少个，这个“滚雪球”式的方法都适用——不断将两个正方形剪拼成一个正方形，直到最后只剩一个大正方形为止。因此，我们得到了一个更强的定理。

结论 2 任意多个正方形都能剪拼成一个大正方形。

其实，在结论 2 的证明中，我们不自觉地用到了图形剪拼的一些最基本的性质。下面，我们将不加证明地提出两个显而易见的结论，考虑图形剪拼的实际过程，这两个结论不难理解。

结论 3 如果图形 A 能剪拼成图形 B，图形 B 能剪拼成图形 C，则图形 A 也能剪拼成图形 C。

结论 4 如果图形 A 能剪拼成图形 B，则图形 B 也能剪拼成图形 A。

利用这两个基本结论，我们能够推出很多有意思的东西。比方说，定义一个“棋盘”是由一个个大小相同的小方格相连形成的图形。由于任意形状的棋盘都能切分成若干个小正方形，而任意多个正方形都能剪拼成一个大正方形，于是，利用结论 3 我



们便可知道，不仅仅是本节开头提到的图形，任意形状的棋盘都能够化成一个大正方形（见图 4）。联想到结论 4，我们又可以得到一个看上去更帅的结论：一个正方形能剪拼出任意形状的棋盘。

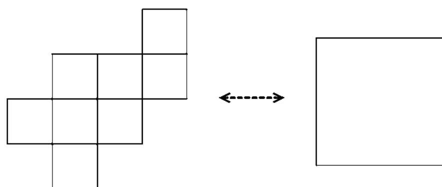


图　4

我们更进一步地思考，那么任意一个矩形是否也总能剪拼成正方形呢？有人可能会说，那还不简单，把这个矩形沿着水平线和竖直线切成一个个的小正方形，然后利用前面的结论不就可以了吗？比方说，把一个长宽比为 3:2 的矩形分成 6 个小正方形，然后再用结论 2 不就能变成一个大正方形了吗？且慢！不见得任意一个矩形都能像这样分成一个个的小正方形。例如，由于 $\sqrt{2}$ 不是有理数，换句话说它不能表示成两个整数之比，因此长宽比为 $\sqrt{2}:1$ 的矩形就不可能划分成一个棋盘。看来，我们还得想想别的招儿。

如图 5 所示，假设矩形 $ABCD$ 的长 $AB=a$，宽 $BC=b$，则与它面积相等的正方形边长就应该是 $\sqrt{ab}$。在 DC 上截取 $DG=\sqrt{ab}$，在 BA 上截取 $BH=\sqrt{ab}$。连接并延长 CH，与 DA 的延长线交于点 E。过 G 做 DC 的垂线，过 E 作 DE 的垂线，两条垂线交于点 F。显然四边形 $DEFG$ 也是一个矩形。

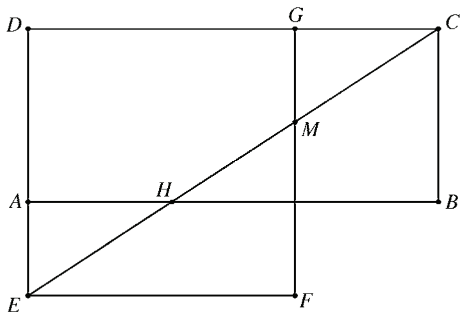


图　5

容易看出，△AHE 和△GCM 全等，△EFM 和△HBC 全等，因此把矩形 $ABCD$ 沿 CH、GM 剪开并分成三块后，可以拼成矩形 $DEFG$。由于这个新矩形的一边长为 $\sqrt{ab}$，而它的面积和原矩形一样都是 ab，可知它的另外一边也是 $\sqrt{ab}$，也就是说它是一个正方形。

但是，这个办法有一个限制条件：矩形的长不能超过宽的 4 倍，否则△GCM 会有一部分跑到矩形外面去，如图 6 所示。

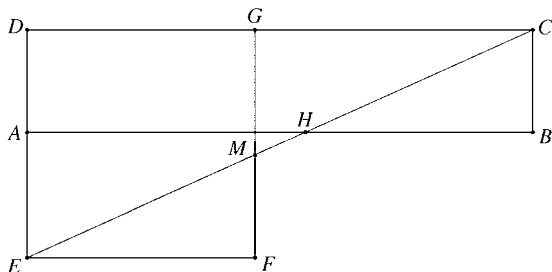


图　6

因此，我们得到了一个并不太完美的结论。

结论 5　若矩形的长宽之比小于 4:1，则这个矩形可以剪拼成一个正方形。

如果矩形的长宽比超过了 4:1 的话，我们还能把它剪拼成一个正方形吗？答案是肯定的。如图 7 所示，如果沿着与较短边平行的线条把一个矩形分成两半，再把其中一块放到另一块的上面，我们就能得到一个新的矩形。这个新矩形的长缩小到了原来的一半，宽却变成了原来的 2 倍，换句话说长与宽的比值减小到了原来的 $\frac{1}{4}$。不断重复这一操作，我们最终总能得到长宽比小于 4:1 的矩形。

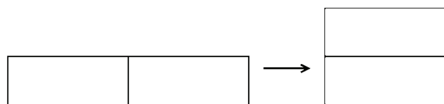


图　7

结论 6　任意一个长宽比超过 4:1 的矩形都可以剪拼成一个长宽比小于 4:1 的矩形。



把上面两个结论综合一下，再利用结论 3，我们就得到了一个完美的结论。

结论 7 任意一个矩形都可以剪拼成一个正方形。

大家可能会很快想到，对于任意平行四边形，我们都可以像图 8 那样，把它变成一个矩形。

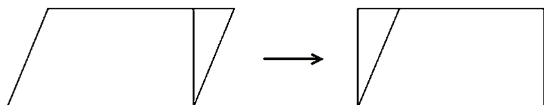


图 8

因此我们有：

结论 8 任意一个平行四边形都可以剪拼成一个矩形。

而刚才说过，任意矩形都能变成正方形。这样，任意一个平行四边形也都能剪拼成一个正方形了。

沿着三角形的中位线将一个三角形分成两块，我们立即发现所有三角形都能变成一个平行四边形（见图 9）。

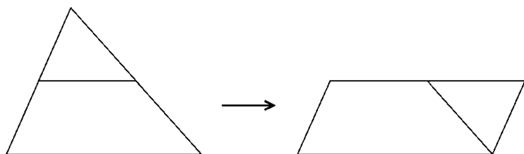


图 9

因而我们有结论 9。

结论 9 任意一个三角形都可以剪拼成一个平行四边形。

这样一来，任意一个三角形也都能够剪拼成正方形了。注意到任意一个四边形总能分成两个三角形，每个三角形都能剪拼成一个正方形，而两个正方形又可以合成一个大正方形，于是我们又可以得到结论 10。

结论 10 任意一个四边形都可以剪拼成一个正方形。

事实上，不仅仅是四边形，任意一个多边形都能分割成若干个三角形，而每个三角形都可以剪拼成一个正方形，任意数目的正方形又能合成一个大正方形（见图 10）。

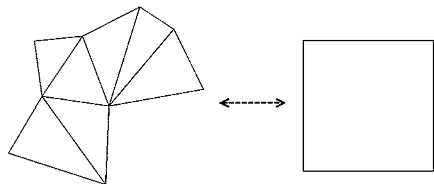


图　10

于是，我们得到了一个更加疯狂的结论。

结论 11　任意一个多边形都可以剪拼成一个正方形。

任意给定两个多边形 A 和 B，由结论 11 可知它们都能剪拼成正方形。如果给定的两个多边形面积相等，那么它们各自剪拼出来的正方形也应该具有相同的边长，换句话说多边形 A 和多边形 B 可以剪拼为同一个正方形。由结论 4 可知，这个正方形也可以反过来剪拼成多边形 A 或者多边形 B。再结合结论 3，我们便得到了这两个多边形之间互相转化的方法：多边形 $A \Leftrightarrow$ 正方形 $\Leftrightarrow$ 多边形 B。于是，我们得到了一个乍看之下很不可思议的神奇定理。

结论 12　任意给定两个面积相等的多边形，它们互相之间都可以通过剪拼得到！

这个定理叫做波尔约–格维也纳定理，是由法卡斯·波尔约（Farkas Bolyai）和保罗·格维也纳（Paul Gerwien）两位数学家分别在 1833 年和 1835 年证明的。



第四部分 精妙的证明

寻找数学之美的过程本身，其实也是一种数学之美。一个漂亮的数学结论，往往有一个漂亮的数学证明。我搜集过很多巧妙到近乎疯狂的证明，每一个证明都会让你拍案叫绝。



26. 我最爱的一个证明

大概是我读高一的时候吧，有一天，我在网上看到了下面这个问题。

设想一个平面上布满间距为 1 的水平直线和竖直直线，形成由一个个单位正方形组成的网格。任意给定一个面积小于 1 的图形，证明这个图形总能放在网格中而不包含任何一个格点。

乍看之下，这简直就是一个世界级的难题，我自然是毫无思路。我滚动鼠标滚轮，继续往下看。出人意料的是，整个证明过程只占据了不到半个屏幕。

我们可以换一个角度来考虑这个问题：把图形随意放在网格中，如何重新布置网格使每个格点都在图形外面。

如图 1 所示，把给定的图形随意放在网格中，然后沿着网格线，把包含有图形的网格切成一个个1×1的小格子，从网格中拿出来。把它们全部重叠起来（不要旋转），再想象这些格子是透明的，而格子上的图形则是不透明的。从上往下看这一叠格子，你看到的会是这个图形的各个部分重叠地放在一个格子中，仿佛一个沾有污渍的方块。由于整个图形的总面积小于 1，因此这些“污渍”不会布满整个方块，方块上总有一块干净的地方。现在，用一根针从一个干净的地方刺下去，把这些重叠起来的方格刺穿。把这些格子放回原来的网格中，你将会看到每一个有图形的方格内都有一个针眼，这些针眼都不在图形内。把原来的网格擦掉，把这几个针眼看作是新网格的格点。按针眼的位置重画网格，那么新网格的所有格点都在原图形之外。这样，结论也就证出了。



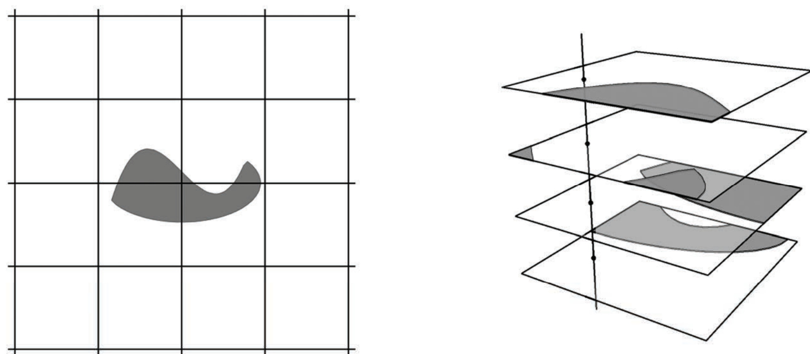


图 1

我还记得当时看完这个证明后，那种无法言表的震撼之感。这个证明太漂亮了！这可能是我第一次如此强烈地体会到数学证明的美妙。之后，我便有意识地去收集各种精彩的数学证明。这些数学证明的思路一个比一个巧妙，方法一个比一个诡异。



27. 把辅助线作到空间中去的平面几何问题

平面几何问题的证明手段无奇不有。有要作辅助线的，有要作辅助圆的，有需要旋转和翻折的……不过，你见过把辅助线作到三维空间中去的问题吗？有时候，一个平面几何问题，放到三维空间中去思考反而会更容易一些。先讲一个我最喜欢的例子吧。

问题 平面上有 4 条直线，任意 2 条不平行，任意 3 条不共点。A、B、C、D 这 4 个人分别在这 4 条直线上匀速行走（他们的速度可以不相同）。若 A 在行走过程中与 B、C、D 相遇，B 在行走过程中与 C、D 相遇（当然也遇见了 A），求证 C、D 在行走过程中相遇。

证明 作垂直于平面的直线作为时间轴，建立三维直角坐标系。由于 4 个人均匀速行走，因此他们的位置–时间图像是线形的。我们可以在空间中作出 A、B、C、D 这 4 个人的位置与时间关系的图像，并分别命名为 l_a、l_b、l_c、l_d。这样，我们就能从这 4 条空间直线中轻易判断某一时刻 4 个人的位置。例如，空间中 P 点 (x, y, t) 在直线 l_c 上，则表明在 t 时刻 C 走到了 (x, y) 的位置。接下来就精彩了。A、B 不是曾经相遇过吗？这就是说，l_a 和 l_b 将会相交。这两条相交直线可以确定一个平面。C 不是与 A、B 都相遇过吗？那就是说，l_c 与 l_a、l_b 都相交。于是，l_c 也在这个平面上。同样地，l_d 也在这个平面上。既然它们全部都共面，l_c、l_d 必然会相交，即 C、D 将会相遇，结论得证。

如果你说，这个问题不算纯粹的平面几何问题的话，那就来看看下面这个问题吧。

问题 如图 1 所示，三角形 ABC 是等边的，P 为三角形内接圆上一点。求证，$AP^2 + BP^2 + CP^2$ 为常数。



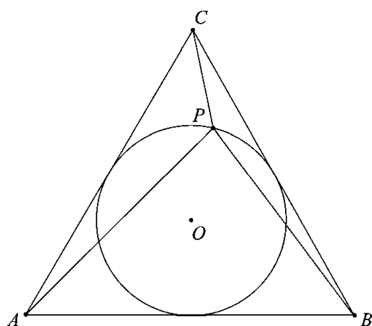


图 1

证明 如图 2 所示，把整个图形放在三维空间里，其中 $A=(1,0,0)$，$B=(0,1,0)$，$C=(0,0,1)$。因此，三角形 ABC 位于平面 $x+y+z=1$ 上。图中的内接圆即为某个以原点为球心的球面 $x^2+y^2+z^2=r$ 与该平面相交所得（其中 r 是某个常数）。

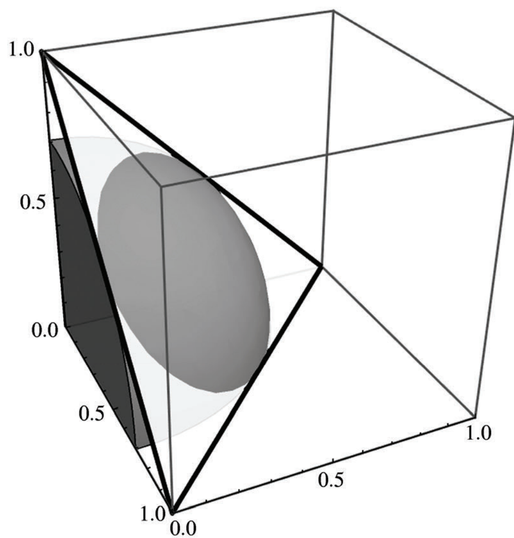


图 2

如果把 P 点的坐标记作 (x_0, y_0, z_0)，由于它既在三角形上又在球面上，因而它将同时满足 $x_0+y_0+z_0=1$ 和 ${x_0}^2+{y_0}^2+{z_0}^2=r$。于是，我们有：

$$
\begin{aligned}
& AP^2+BP^2+CP^2 \\
&=(1-x_0)^2+y_0{}^2+z_0{}^2+x_0{}^2+(1-y_0{}^2)+z_0{}^2+x_0{}^2+y_0{}^2+(1-z_0)^2 \\
&=3\cdot(x_0{}^2+y_0{}^2+z_0{}^2)-2\cdot(x_0+y_0+z_0)+3 \\
&=3\cdot r-2+3
\end{aligned}
$$

结果是一个常数。

把平面问题扩展到三维空间，也不都只是为了借用空间直角坐标系这一工具。空间几何中的一些已知结论，也能在平面几何问题中派上用场。

问题　如图 3 所示，考虑平面上的一个任意三角形 *ABC*，以及异于 *A*、*B*、*C* 三点的一个点 *P*。*X*、*Y*、*Z* 分别是 *P* 点关于 *BC*、*AC*、*AB* 三边的中点的对称点。求证：*AX*、*BY*、*CZ* 共点。

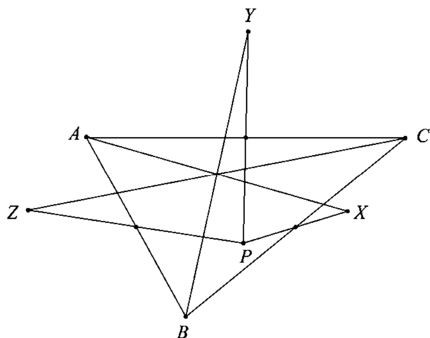


图　3

证明　如图 4 所示，考虑空间中的一点 *P′*使得 *PP′*垂直于平面 *ABC*。作出 *P′*关于 *BC*、*AC*、*AB* 三边中点的对称点 *X′*、*Y′*、*Z′*。容易看出，四边形 *P′AY′C*、*P′BZ′A* 和 *P′CX′B* 都是平行四边形。那么，*A*、*B*、*C*、*P′*、*X′*、*Y′*、*Z′*就成了一个平行六面体的其中七个顶点。*AX′*、*BY′*、*CZ′*是平行六面体的三条体对角线，它们是共点的。现在，把 *P′*、*X′*、*Y′*、*Z′*全部投影到平面 *ABC* 上，就是我们想要的结论了。



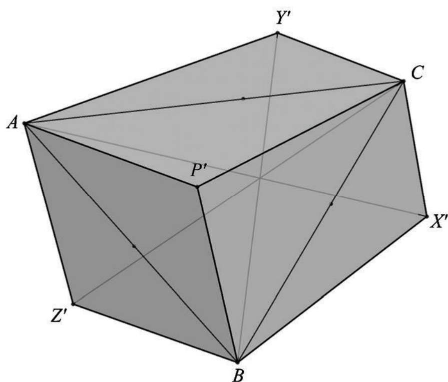


图 4

一些非常经典的初等几何定理也有让人意想不到的“空间解法”。

问题 如图 5 所示，在△ABC 中，点 D、E、F 分别在 BC、AC、AB 所在直线上，若 D、E、F 三点共线，则有 $\frac{AF}{BF}\cdot\frac{BD}{CD}\cdot\frac{CE}{AE}=1$。

证明 如图 6 所示，过 DEF 所在直线作一个新的平面。分别过 A、B、C 作原平面的垂线，与新的平面交于点 A'、B'、C'。由相似三角形的关系不难看出，$\frac{AA'}{BB'}=\frac{AF}{BF}$，并且 $\frac{BB'}{CC'}=\frac{BD}{CD}$，另外还有 $\frac{CC'}{AA'}=\frac{CE}{AE}$。三个等式乘在一块儿，结论得证。

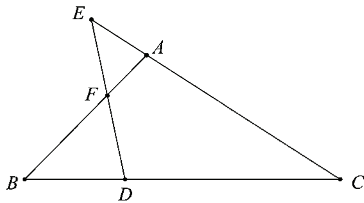


图 5

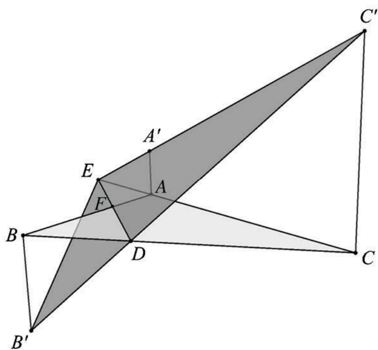


图 6

上面这个定理叫做梅涅劳斯（Menelaus）定理，是平面几何中一个非常重要的定

理，我们常常用它来判断三点共线。梅涅劳斯定理的证明方法有很多，上面这种方法恐怕是最帅的一种了。它解决了其他证明方法缺乏对称性的问题，完美展示了几何命题中的对称之美。

下面则是另一个经典的几何定理，它叫做蒙日定理，是由法国数学家加斯帕德·蒙日（Gaspard Monge）首次发现的。如果是第一次看到这个定理，你一定会被它深深地迷住；而它的空间证明方法，则更是叫人为之倾倒。

问题 如图7所示，平面上有三个互相分离的圆，其中任意两个圆都有两条外公切线，这两条外公切线交于一点。显然，这样的交点共有三个。求证，这三点共线。

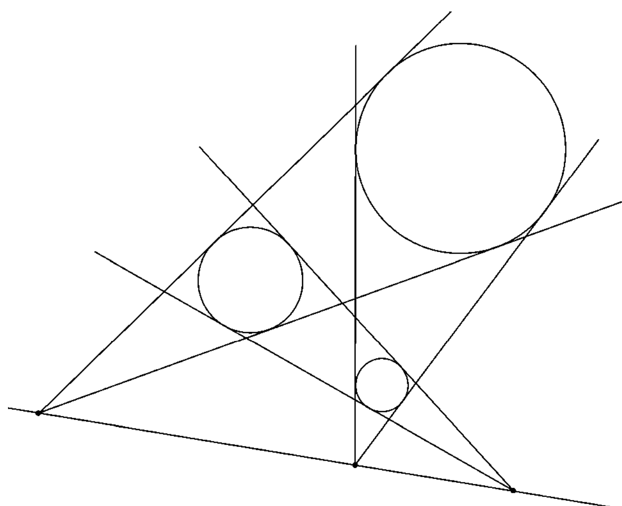


图 7

证明 如图8所示，在这个平面的三个圆上放置三个球，每个球的半径都等于它底下的那个圆的半径。显然，这个平面是这三个球的一个公切面。再把三组公切线想象成这三个球两两确定的三个圆锥在平面上的投影。显然，三个圆锥的顶点都在这个平面上，我们要证明的就是，这三个顶点是共线的。注意到这三个球还有另一个公切面（想象一块薄玻璃板从上面盖下去），三个圆锥的顶点也都在这个公切面上。而这两个公切面的公共部分就是它们的交线，因此三个顶点必然都在这条交线上。



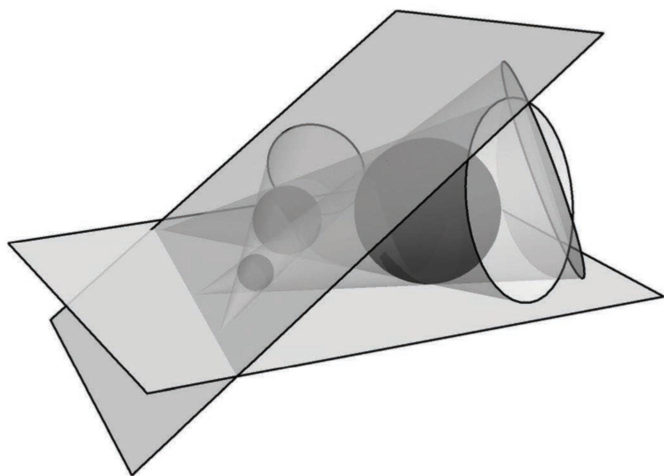


图 8

最后让我们来看一个例子。下面这个平面几何问题乍看之下也很难解决，但借助“辅助球”的思想，结论变得几乎是平凡的了。

问题 平面上三个圆两两相交。试证明三条公共弦共点（见图9）。

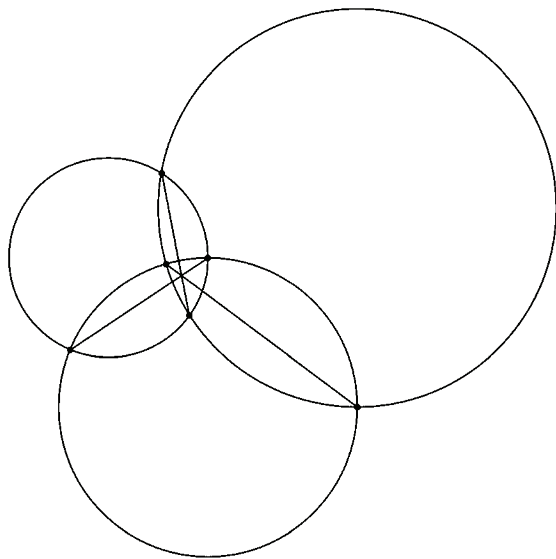


图 9

证明　分别以这三个圆为“赤道面”，作出三个球体。我们把这三个球的球心（也就是原问题中的三个圆心）所确定的平面（也就是原问题的图形所在的平面）记作 α。注意到，每两个球面将会相交于一个圆圈，它们在 α 上的投影就是那三条公共弦。而三个球面将会交于两个点（这两个点一上一下，关于 α 对称），并且这两个点都同时属于空间中的三个圆圈。从投影的角度来看，这就是说，在平面 α 上存在一个点，它同时属于那三条公共弦。这就说明了三条公共弦交于一点。

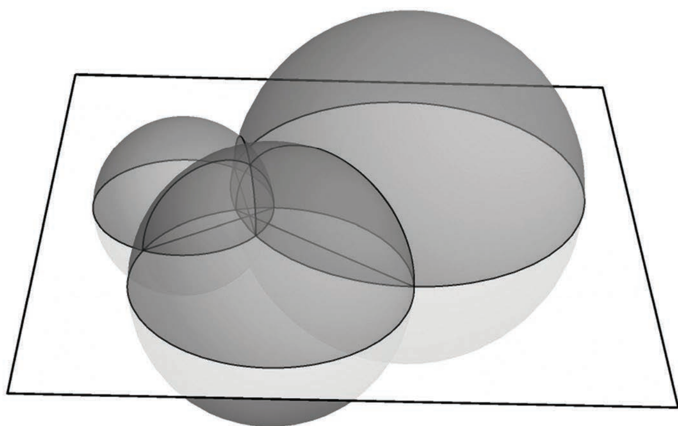


图　10



28. 小合集（一）：几何问题

有一次，我临时帮忙去代一节小学奥数课，见到了下面这道题。

如图 1 所示，*ABCD* 是一个正方形，边长为 4，*DEFG* 是一个矩形，其中 *DG*=5，求 *DE* 的长度。

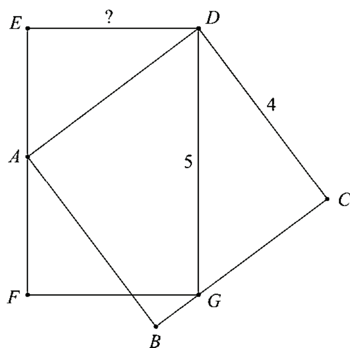


图 1

当然题目本身并不难，或许有的人一眼就能看出答案。问题的关键在于，这个问题是一道小学奥数题，这意味着这个题目一定有一个异常巧妙的傻瓜解——不能列方程，不能用相似形，事实上几乎什么都能不用，只需要用到最基本最显然的正方形长方形的性质。

我叫了几个初中数学老师来，一起围着它研究了半天，结果想破脑袋也还是满脑子的相似三角形，于是只好求助小学组的老师，果然取得真经，赞不绝口，大呼妙哉。

如图 2 所示，连接 *AG*。注意到三角形 *ADG* 的面积既是正方形 *ABCD* 面积的一半，又是矩形 *DEFG* 面积的一半，可见正方形和矩形的面积是相等的。既然正方形的面积是 16，矩形的一边长是 5，另一边就是 3.2 了。

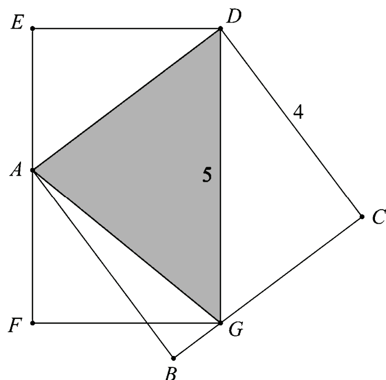


图　2

这让我立即想到了下面这个题目：

如图 3 所示，其中阴影部分是一个正方形，求该正方形的边长。

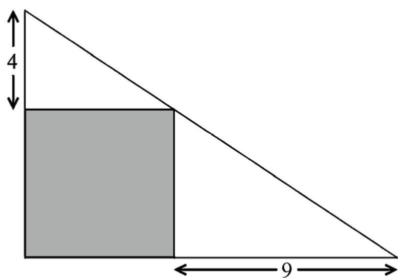


图　3

这也是一道小学奥数题。一个小学奥数老师曾经告诉我，当年带领学生参加这次竞赛时，领队老师们都没有想到"小学生解法"，以至于开始质疑这道题是否超纲了。看到答案后，老师们大为折服——这个问题确实有一个无需任何几何知识的妙解。

把图形补充为一个长方形（见图 4），则两个大的直角三角形面积相同，另外还有 *A* 的面积与 *B* 的面积相同，*C* 的面积与 *D* 的面积相同。于是我们得到，阴影部分与右上角的那个小长方形面积相同，而后者的面积应该是 36。这就是说，正方形的边长应该等于 6。



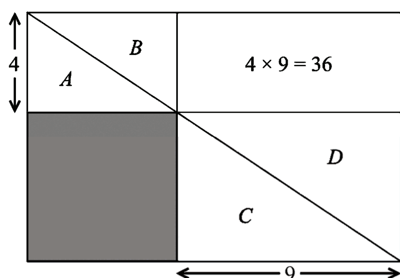


图 4

千万别小看这种雕虫小技。利用这种面积相等的关系，我们能证明不少经典的定理，下面就是其中一个定理。

在平面直角坐标系中有 (a,b) 、(c,d) 两点（见图 5）。为了简便起见，不妨假设它们都在第一象限。将这两个点分别与原点相连，然后以这两条连线为边，作一个平行四边形。求证：这个平行四边形的面积为 $|ad-bc|$ 。

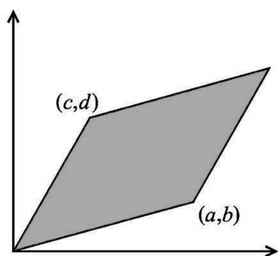


图 5

只需简单的三步，结论就显而易见了。

第一步，把平行四边形上面突出的部分平移下来（见图 6）。第二步，把左右突出的部分也都平移到对面去。这样一来，整个图形就变成了一个工工整整的矩形。第三步，也是最关键的一步，把图中那块阴影矩形移到左上方一块与其面积相等的空白处（这里用到了上一个题目里的等积关系）。这下就再清楚不过了，整个图形的面积就是 $ad-bc$。

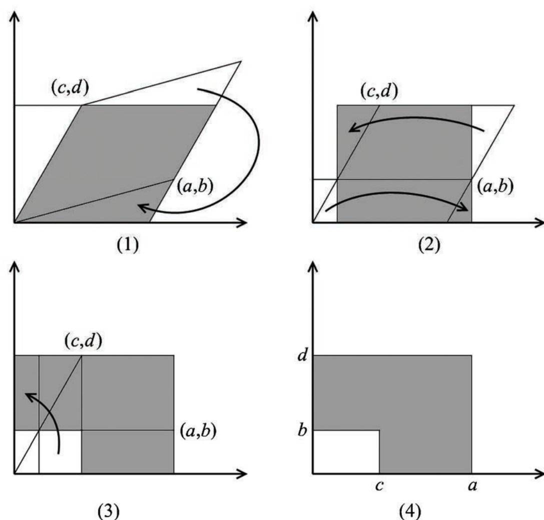


图　6

注意，如果 (a,b) 和 (c,d) 的位置互换一下，得到的面积值将会是 $bc-ad$ ，和前面的计算结果互为相反数。因此，把平行四边形的面积记作 $|ad-bc|$ 要更好一些。

另外，注意看由原点、(a,b) 和 (c,d) 所组成的三角形的面积，它应该是平行四边形面积的一半，也就是 $\frac{|ad-bc|}{2}$。在下一节中，我们将会用到这种三角形面积计算法。其实，这就是向量叉积的性质，而我们仅仅用小学知识就证明了它！

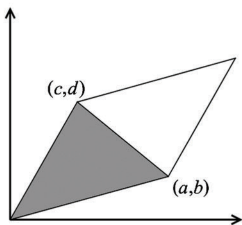


图　7

在教初中数学竞赛时，我也遇到过一些类似的趣题妙解。其中一个题目如下：

五个圆依次相切，它们又都相切于两条不平行的直线（见图 8）。如果最左边那个圆的半径为 4，最右边那个圆的半径为 9，求中间那个圆的半径。



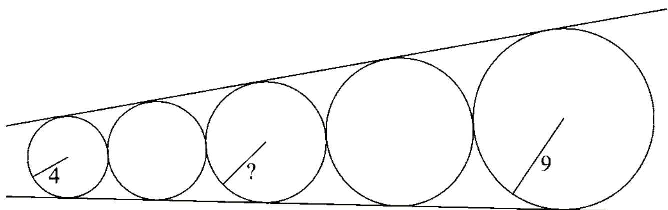


图 8

有趣的是，这也有一个非常直观的解答方法：

中间那个圆的半径为 6。下面我们说明，事实上五个圆的半径是成等比数列的。把五个圆从小到大依次记作 C_1、C_2、C_3、C_4、C_5，把两条直线的交点记为 P。把 C_1、C_2 的圆心到 P 的距离分别记作 P_1、P_2。现在，把整个图以 P 为中心缩小到原来的 $\frac{P_1}{P_2}$，则两条直线还在原来的位置，但是现在 C_2 的圆心恰好和原来 C_1 的圆心重合。另外，由于缩放不会影响相切关系，因此现在的 C_2 就和原来 C_1 完全重合，同理现在的 C_3 就是原来的 C_2，现在的 C_4 就是原来的 C_3，现在的 C_5 就是原来的 C_4。另外，由于整个图形都缩小到了原来的 $\frac{P_1}{P_2}$，因此每个圆也都缩小到了原来的 $\frac{P_1}{P_2}$，而此时每个 C_i 正好和原来的 C_{i-1} 一样大了。这就说明，每两个相邻圆的半径之比为 $\frac{P_1}{P_2}$。

下面则是一道非常帅气的高中数学竞赛题，它是我从好友范翔[①]那儿听来的。

求证：当 n 为奇数时，用 $n-3$ 条对角线将正 n 边形分为 $n-2$ 个三角形，则有且仅有一个三角形是锐角三角形（图 9 仅仅画了 $n=9$ 时的其中一种分割方案）。

据说，在一堂高中奥数课上，老师出了这道题后等了半个小时，大家被搞得焦头烂额，也没有想出半点思路。这时，老师开始讲题了。只一句话，所有人都恍然大悟了，然后集体开始鼓掌——这个证明实在是太巧妙了！证明会用到这么一个定理：一个三角形是锐角三角形，当且仅当它的外心在这个三角形的内部。

---

① 范翔的博客地址为 http://www.eaglefantasy.com。

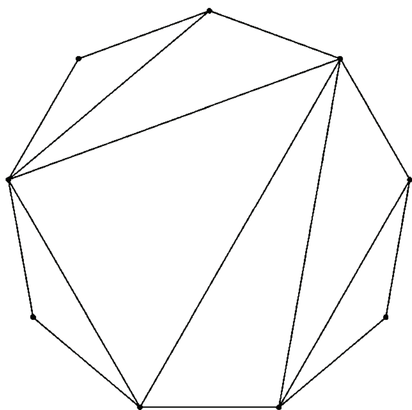


图　9

作出正 n 边形的外接圆，这个外接圆的圆心也就成了所有三角形公共的外心。这个外心一定位于某个三角形的内部，它就是唯一的那个锐角三角形。

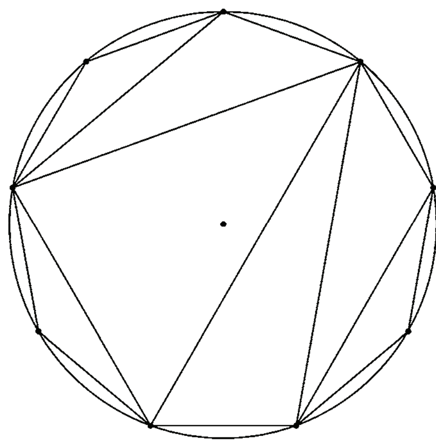


图　10

这让我联想到另一个有异曲同工之妙的问题：椭圆内能否内接一个正五边形（假如圆不算椭圆）？答案是否定的。否则，作出这个正五边形的外接圆，那么它将会和椭圆有五个交点，而这显然是不可能的。

1979 年北美普特南（Putnam）数学竞赛中的 A4 小题则是又一道让人拍案叫绝的精彩好题。



平面上有 n 个红点和 n 个蓝点，其中任意三点不共线。你需要把它们一红一蓝地配成 n 对，并用线段把每一对点连接起来。证明，总存在一种配对方案，使得所有连线都不交叉。

这个问题看起来似乎相当困难，其实整个证明就只有几句话。

考虑所有可能的配对方案，选择所有连线的长度总和最小的那一种方案。下面我们证明，这种方案是满足要求的。如图 11，假如在这种方案中有某四个点 A、B、C、D，其中红点 A 和蓝点 B 相连，红点 C 和蓝点 D 相连，两条连线交于点 O。那么，把它们改成 A 与 D 相连，B 与 C 相连，则由三角形两边之和大于第三边知，$AB+CD=(AO+DO)+(BO+CO)>AD+BC$，这说明连线的总长度变得更短了，由此产生矛盾。

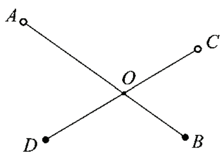


图 11

下面这个题目则来自 2011 年 IMO 国际奥林匹克数学竞赛中的第二题。

假设平面上有若干个点（至少两个），其中任意三点都不在同一条直线上。所谓一个“风车”是指这样一个过程：从经过某一点 P 的一条直线 l 开始，以 P 为中心顺时针旋转，直到这条直线碰到另一个点，比如说点 Q。接着将这条直线以 Q 为新的旋转中心进行顺时针旋转，直到再次碰到其他的点，并像这样一直持续下去。证明，我们总可以适当选取某一点 P，以及过 P 的一条直线 l，使得由此产生的“风车”最终将会碰到所有的点。

比方说，考虑一个等边三角形的三个顶点和这个三角形的中心所组成的四个点。从图 12 左边所示的直线出发，就能碰到所有的点；但从图 12 右边所示的直线出发，只能围绕三角形的三个顶点转下去。

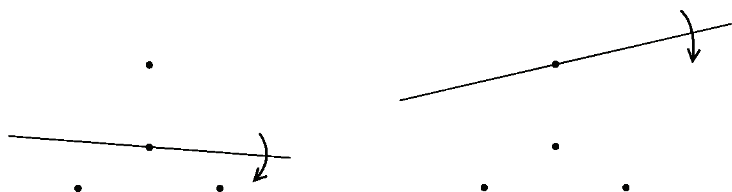


图　12

赛后统计资料显示，这道漂亮的问题竟然是六道题中第二难的题。著名的数学博客 polymath blog 在网上组织了 mini-polymath3 活动，邀请众人一同讨论这道题的解法。活动一开始，便引来各路数学高人献计献策，提出了很多有趣的思路和猜想。第 74 分钟，终于有人给出了正确的解法。果然不出所料，神题就该有神解，这道题有一个异常简单巧妙的证明方法。

如图 13 所示，找一条直线，这条直线两侧的点数一样多（最多相差一个）。下面我们证明，这条直线就满足要求。容易看出，在直线的旋转过程中，直线两侧的点数之差始终不变。因此，这条直线转了 180 度后，一定回到了初始的位置（或者它旁边一个点的位置）。但此时，原来在直线左侧的点现在全部跑到了直线右侧，原来在直线右侧的点现在全部跑到了直线左侧。这些点当然是不能“瞬移”到直线另一侧的，要想跑到直线的另一侧，必须要先穿过直线才行。由此可见，所有点都被直线碰到过了。

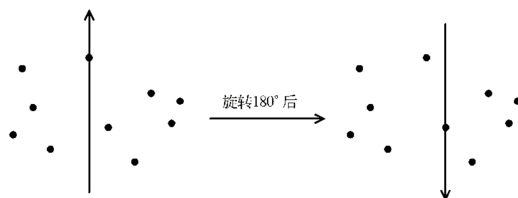


图　13

别以为这种趣题妙解只会出现在竞赛题目中。在真正的数学研究过程中，也常常出现让人叹为观止的简短证明。1941 年，数学家保罗 · 埃尔德什在《美国数学月刊》上提出了下面这个问题。



如图 14 所示，在一个单位正方形内，有两个互不重合的小正方形。求证，这两个小正方形的边长之和不可能大于 1。

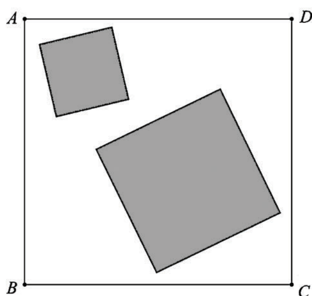


图 14

下面则是一个极其巧妙的证明。

在证明过程中，我们会用到这么一个定理：在一个直角三角形内画一个正方形，则图 15 所示的正方形具有最大的面积。这可以用分情况讨论的办法进行证明，只是计算过程有些复杂，这里就不再详述了。

如图 16，如果两个正方形是完全分离的，那么一定能找出一条从它们中间穿过去的直线 *XY*。假设它和另一个方向上的对角线相交于 *P*。从 *P* 点出发，向大正方形的四条边分别作垂线。于是，直线 *XY* 上方那个小正方形的面积不会超过正方形 *AMPN* 的面积，直线 *XY* 下方那个小正方形的面积不会超过正方形 *PSCT* 的面积。这就告诉我们，上面那个正方形的边长不超过 *AN*，下面那个正方形的边长不超过 *SC*，也即两个正方形的边长和不超过 1。

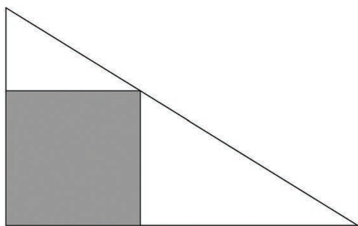


图 15

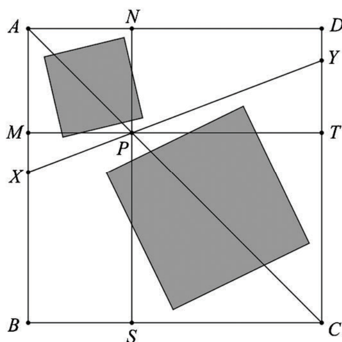


图 16

最后，让我们来看一个更富戏剧性的例子。你能不能在纸上画一些点，使得任意两点所确定的直线都会经过第三个点？当然，所有点都在一条直线上的情况除外。英国数学家詹姆斯·约瑟夫·西尔维斯特（James Joseph Sylvester）认为不可能。1893年，他提出了下面这个猜想。

若 n 个点不全共线，则必存在一条直线恰好穿过两个点。

不过，西尔维斯特却不能证明这一点。直到 1933 年，数学家蒂伯·加莱（Tibor Gallai）才给出了一个证明，不过证明过程相当复杂。1948 年，戏剧性的一幕终于发生了：保罗·约瑟夫·凯利（Paul Joseph Kelly）发现，这个结论竟然有一个简单得令人瞠目结舌的证明。

假设存在某个点集，任意两点确定的直线上都存在其他的点。画出所有可能的直线，作出每一个点到每一条直线的垂线段，然后找出这些垂线段中最短的一条。不妨假设这条最短的垂线段是点 P 到某条直线 l 的垂线段，垂足点记作 H。由假设，l 上至少有三个点，因此至少有两个点分布在垂足 H 的同一侧（允许和垂足重合）。不妨把这两个点记作 R、Q，如图 17 所示。由于我们画出了所有可能的直线，因此 P、R 两点之间也有一条直线；此时，Q 到 PR 的垂线段就是更短的垂线段，于是产生矛盾。要想避免这样的矛盾，唯一的方法就是，所有的垂线段长度都为 0，换句话说根本作不出所谓的垂线段。这也就是 n 个点全都共线的情况。

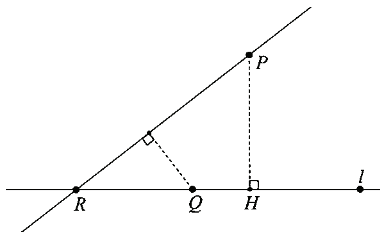


图 17

这个证明思路妙到了极点，几十年来愣是没有一个人发现！



29. 皮克定理的另类证法和出人意料的应用

图 1 中的每个小正方形面积都是 1，那么图中的三角形面积是多少？

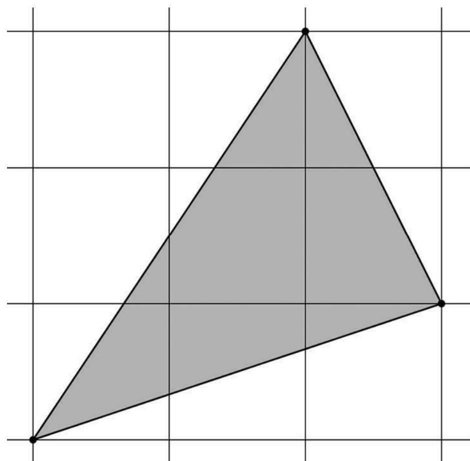


图 1

你会发现，传统的三角形面积计算公式（底乘以高除以 2）在这里已经不管用了。三角形的三边长度都带有根号，高的长度更难求出。计算三角形面积还有一个常用的公式，叫做海伦（Heron）公式：如果三角形的三边长分别为 a 、b 、c ，周长的一半为 s ，则其面积等于 $\sqrt{s(s-a)(s-b)(s-c)}$ 。用这种方法倒是能算出三角形的面积，不过运算过程过于复杂，并不可取。能否把三角形剪剪拼拼，割补成一个更规则的图形呢？多试几下，好像也是办不到的。种种失败似乎暗示着，这个三角形的面积并不那么好求，它恐怕是一个非常复杂的代数式吧。

其实，这个三角形的面积值是一个非常简单的数——3.5。还记得上一节讲到的三

角形面积叉积计算法吗？把三角形中左下角的那个顶点当作原点，套用公式我们可以立即算出，三角形的面积就是 $\frac{3\times3-1\times2}{2}=\frac{7}{2}$，也就是 3.5。其实，我觉得，连这种方法都复杂了。稍稍换一个角度，我们还有一个简单得你都不敢相信的算法。

如图 2，整个三角形完全包含于一个面积为 9 的大正方形内。减去直角三角形 A 的面积 3，减去直角三角形 B 的面积 1.5，再减去直角三角形 C 的面积 1，就得到了我们要求的三角形面积，它等于 3.5。

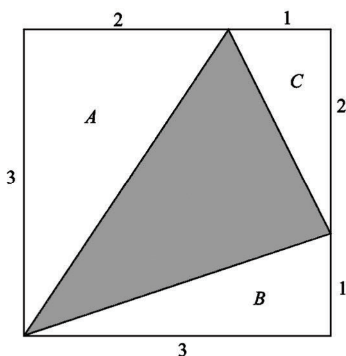


图　2

1899 年，奥地利数学家乔治·亚历山大·皮克（Georg Alexander Pick）发现，不止是三角形，对于平面上的任意一个多边形，只要它的每个顶点都在单位正方形网格的“格点”上，它的面积都有类似的巧算方法。皮克沿着这个思路进一步推导，得出一个超级简单的面积计算公式：令 I 等于多边形内部所含的格点数，令 B 等于多边形边界上的格点数，则多边形的面积就是 $I+\frac{B}{2}-1$。这就叫做皮克定理。

在这里，我们舍弃复杂的分类讨论和数学归纳法，用一种全新的方法来证明这个结论。假设整个平面是一个无穷大的铁板，每个格点上都有一个单位的热量。经过无穷长时间的传导后，最终这些热量将以单位密度均匀地分布在整个铁板上。平面上某个多边形内所含的热量，也就代表它的面积了。下面我们试着求出多边形内的热量。考虑多边形的任意一条线段，如图 3 所示（有觉得图 3 中的这条线段不直吗？那是你的错觉）。由于它的两个端点均在格点上，因此整个平面网格一定关于这条线段的中点对称，因而流经该线段的热量也就是对称的，这半边流出去多少，那半边就流进来



多少，出入该线段的热量总和实际为 0。我们立即看到，多边形的热量其实完全来自于它内部的 I 个格点（的全部热量），以及边界上的 B 个格点（各自在某一角度范围内传来的热量）。边界上的 B 个点形成了一个内角和为 $(B-2)\times180°$ 的 B 边形。这 B 个点本来蕴含了 B 个单位的热量，但只有其中的 $\frac{(B-2)\times180°}{B\times360°}$ 这一比例的热量流入了多边形。因此，从这 B 个点流入多边形的热量就等于 $B\cdot\frac{(B-2)\times180°}{B\times360°}=\frac{B-2}{2}=\frac{B}{2}-1$。再加上 I 个内部格点的全部热量，于是得到多边形内的总热量（也就是它的面积）就是 $I+\frac{B}{2}-1$。

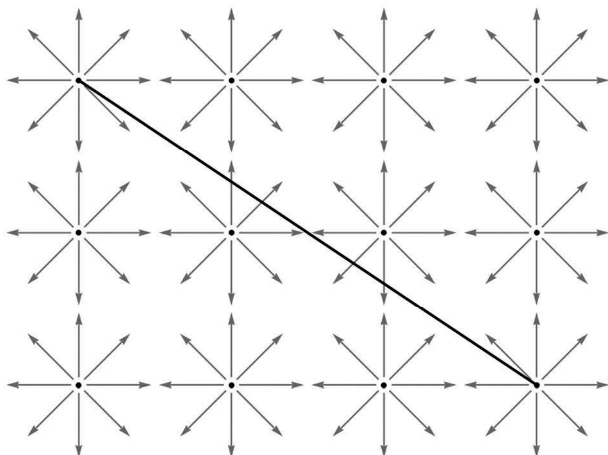


图　3

计算“格点多边形”的面积就是这么简单，不信的话，让我们来试试。例如，图 4 中的第一个多边形，它的内部没有点，边界上有 4 个点，因此面积就是 $0+\frac{4}{2}-1=1$。第二个多边形的形状虽然不一样，但内部也没有点，边界也经过了 4 个点，因此面积也是 $0+\frac{4}{2}-1=1$。第三个图形的边界上也有 4 个点，但内部包含了一个点，因此其面积就是 $1+\frac{4}{2}-1=2$。回头看看本节最早画出的那个三角形，面积就是 $3+\frac{3}{2}-1$，也就是 3.5 了。需要注意的是，皮克定理只适用于所有顶点都在格点上的多边形，其他情况是不能套用皮克定理的。

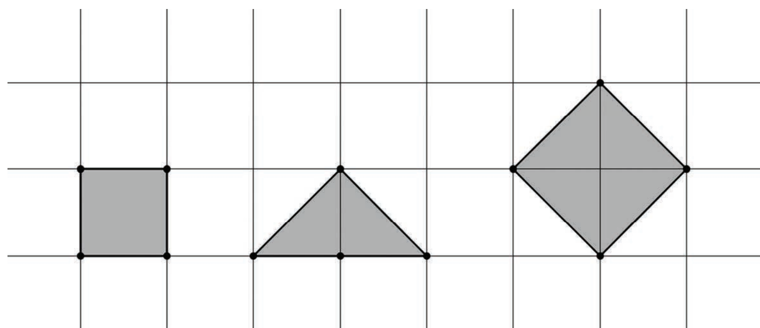


图　4

皮克定理有很多有趣的推论。例如，由皮克定理立即可知，格点多边形的面积一定是$\frac{1}{2}$的整数倍。而这又可以推出，等边三角形的三个顶点不可能都在格点上，也就是说你永远不可能找出三个格点，它们恰好组成一个等边三角形。这是因为，假如等边三角形的边长为s，可以求出其面积为$\frac{\sqrt{3}}{4}s^2$，但由勾股定理可知，两个格点间的连线长度的平方一定是一个整数，即s^2一定是整数，从而$\frac{\sqrt{3}}{4}s^2$一定是无理数，这与格点多边形的面积是$\frac{1}{2}$的整数倍相矛盾。

另一个有趣的推论是，在一个$m \times n$的点阵中画一条经过所有点恰好一次的回路，得到的多边形面积一定是相同的。举例来说，图 5 中的三个多边形，哪一个面积最大？利用皮克定理便能立即看出，它们是一样大的。因为它们都是4×4点阵中的格点多边形，并且所有 16 个格点都用在了多边形边界上，内部显然不可能再有格点了，所以它们的面积都是$0+\frac{16}{2}-1=7$。

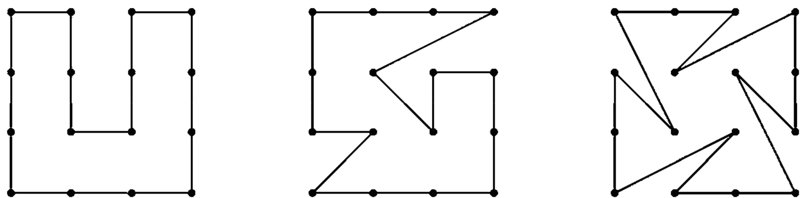


图　5



皮克定理还有一些更精彩的推论。考虑平面直角坐标系中的直线 $x+y=n$，其中 n 是一个质数。这条直线将恰好通过第一象限里的 $n-1$ 个格点（如图 6，图中所示的是 $n=11$ 的情况）。将这 $n-1$ 个点分别和原点相连，于是得到了 $n-2$ 个灰色的三角形。仔细数数每个三角形内部的格点数，你会发现一个惊人的事实：每个三角形内部所含的格点数都是一样多的。这是为什么呢？

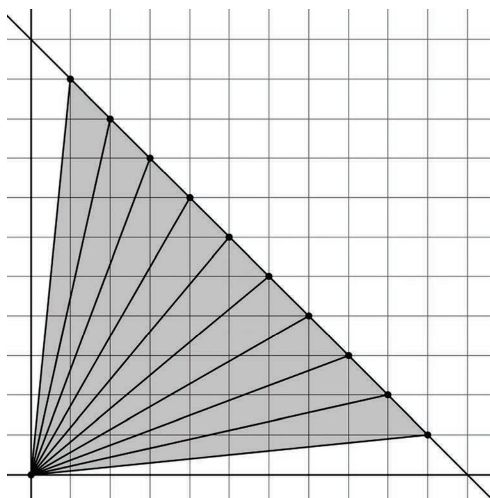


图　6

借助皮克定理，我们能得到一个漂亮的解释。首先，假如正整数 x 和 y 有一个公因数 d，也就是说 x 和 y 都是 d 的倍数，那么显然 $x+y$ 就也是 d 的倍数。反过来，如果 $x+y$ 是一个质数，这就表明 x 和 y 不可能有公因数，换句话说 x 和 y 是互质的。也就是说，点 (x, y) 和原点的连线不会经过其他格点。因此，所有灰色三角形边界上都只有 3 个格点（即三角形的三个顶点），不会再经过其他格点。另外，注意到所有灰色三角形都是等底等高的，因此它们的面积都相等。既然所有三角形的面积都相等，边界上的格点数也相等，由皮克定理可知，每个三角形内部的格点数也都相等了。

一个东西最出神入化的运用还是见于那些本来与它毫不相干的地方。在数论中，法里（Farey）序列是指把 0 到 1 之间的所有分母不超过 n 的最简分数从小到大排列起来所形成的数列，我们把它记作 F_n。例如，F_5 就是

$$\frac{0}{1}, \frac{1}{5}, \frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{1}{1}$$



法里序列有一个神奇的性质：前一项的分母乘以后一项的分子，一定比前一项的分子与后一项的分母之积大 1。更不可思议的是，这竟然可以用皮克定理来解释！

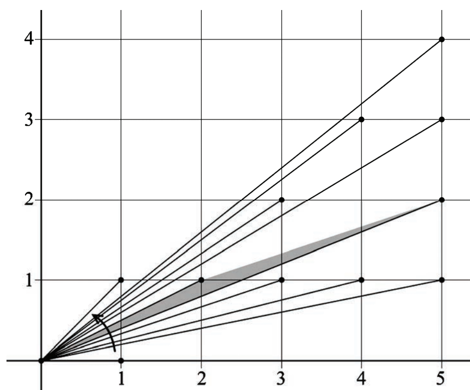


图　7

把每一个介于 0 和 1 之间并且分母不超过 n 的最简分数都标记在平面直角坐标系上，例如 $\frac{0}{1}$ 就对应点 $(1,0)$ ，$\frac{1}{5}$ 就对应点 $(5,1)$ 。一个分数的大小，也就直观地反映为对应的标记点与原点连线的倾斜程度：倾斜程度越小，分数值越小；倾斜程度越大，分数值也就越大。现在，考虑一条以原点为端点的射线从 x 轴正方向出发逆时针慢慢转动到 y 轴正方向，这条射线依次扫过的标记点正好就是一个法里序列。考虑这根射线扫过的两个相邻的标记点，它们与原点所组成的三角形面积是多少呢？我们试着用皮克定理来计算。由于所有分数都是最简分数，也就是说它们的对应点的横纵坐标是互质的，因此它们与原点的连线上没有其他格点；又因为这是射线扫过的两个相邻标记点，因此三角形内部以及这两个标记点的连线上也都没有任何格点。可见，除了边界上的三个顶点外，三角形上再无其他格点了。因此，射线扫过的两个相邻点，与原点组成的三角形面积一定是 $0+\frac{3}{2}-1=\frac{1}{2}$ 。另外别忘了，上一节还讲到了三角形面积的叉积计算法，(a,b) 和 (c,d) 两个点与原点组成的三角形面积应该为 $\frac{ad-bc}{2}$ 。于是，对于法里序列的两个相邻分数 $\frac{b}{a}$ 和 $\frac{d}{c}$ ，我们有 $\frac{ad-bc}{2}=\frac{1}{2}$ ，即 $ad-bc=1$ 。

我们竟然用几何手段，证明了一个与几何毫无关系的数论定理！别吃惊，稍后大家还会看到一个几何定理的数论证法。



30. 欧拉公式的另类证法和出人意料的应用

图 1 中的大格点多边形是由若干个小格点多边形拼成的,怎样计算它的面积呢？我们有两种不同的计算方案。第一种方案是先算出所有小多边形的面积和。如果把所有小多边形的内部格点数之和记作 ΣI ，把所有小多边形各自边界上的格点数之和记作 ΣB ，把小多边形的个数记作 F ，由皮克定理，这些小多边形的面积总和就是 $\Sigma I+\frac{\Sigma B}{2}-F$ 。需要注意的是，ΣB 当中有很多格点都被算了多次。我们通常把一个交叉点的"分岔数"（也就是这个点向外发射的线条数）叫做这个点的"度数"，那么整个大多边形内部的每个顶点被重复计算的次数就等于它的度数,边界上的顶点被重复计算的次数则等于它的度数减 1。而对于线段上的格点，内部线段经过的格点都被重复算过两次，而大多边形边界上的线段经过的格点则只被算过一次。如果我们把内部顶点个数记作 V_i ，把它们的度数之和记作 $\sum D_i$ ，再把边界顶点个数记作 V_b ，把它们的度数和记作 $\sum D_b$ ，最后把所有内部线段经过的格点数之和记作 $\sum S_i$ ，边界上的线段经过的格点数之和记作 $\sum S_b$ ，那么 $\Sigma I+\frac{\Sigma B}{2}-F$ 就可以重新写成 $\Sigma I+\frac{\sum D_i+\sum D_b-V_b+2\times\sum S_i+\sum S_b}{2}-F$ 。

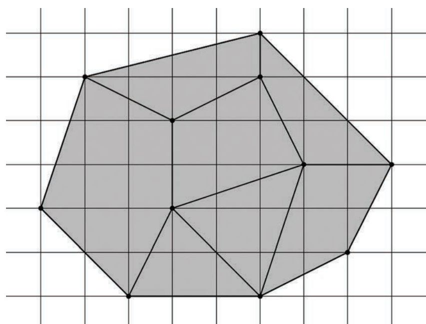


图 1

第二种方案是用皮克定理直接计算整个大多边形的面积，它应该等于 $\Sigma I+\Sigma S_i+V_i+\frac{\sum S_b+V_b}{2}-1$。由于两个式子计算的都是整个大多边形的面积，因此它们的值应该是相等的，也就是说：

$$\Sigma I+\frac{\sum D_i+\sum D_b-V_b+2\times\sum S_i+\sum S_b}{2}-F=\Sigma I+\Sigma S_i+V_i+\frac{\sum S_b+V_b}{2}-1$$

去掉等号两边相同的部分，整个等式可以化简为：

$$\frac{\sum D_i+\sum D_b-V_b}{2}-F=V_i+\frac{V_b}{2}-1$$

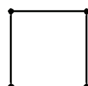
再整理一下，就成了：

$$\frac{\sum D_i+\sum D_b}{2}-F=V_i+\frac{V_b+V_b}{2}-1$$

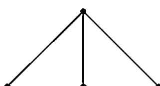
也就是 $\frac{\sum D_i+\sum D_b}{2}-F=V_i+V_b-1$。如果我们把这个图中的顶点总数记作 V，那么显然 V_i+V_b 就等于 V；如果再把这个图中的线段总数记作 E，由于每根线段都贡献了两个度数，因此 E 就等于所有顶点度数和的一半，也就是 $\frac{\sum D_i+\sum D_b}{2}$。至此，我们得到了一个重要的式子：$E-F=V-1$。对于所有本节开头给出的那种图，这个式子总是成立的。

习惯上，我们一般把它写成 $V-E+F=1$。另外，我们一般不说 F 是图中小多边形的个数，而是把它定义为一个更宽泛、更常见概念：图形的“区域数”。如果把整个图形外边无限大的空间也算作一个区域的话，等式就将改写为 $V-E+F=2$。

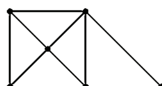
这便是著名的欧拉公式：任意一个图中的顶点数 V、线条数 E 和这些线条围成的区域数 F（包括最外边那个无限大的区域）一定满足 $V-E+F=2$。图 2 给出了一些具体的例子。有两点需要注意：第一，这个图必须是一个完整的连通的图，不能是由好几个分离的小图组成；第二，图中的连线是不能交叉的，如果相交了，交叉点就要算作新的顶点。



$V=4$
$E=4$
$F=2$



$V=4$
$E=5$
$F=3$



$V=6$
$E=10$
$F=6$

图　2



容易想到，即使顶点不在格点上，顶点数、边数和区域数的关系仍然是不变的。事实上，即使点与点之间的线条不是直的，顶点数、边数和区域数的关系依旧不变。更神的是，当我们放开限制后，图中完全有可能出现“两边形”，甚至是“一边形”，即使这样也丝毫不会影响$V-E+F=2$的正确性。有趣的是，这个加强版的结论反而有一个更简单的证明。

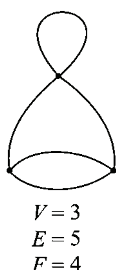


图 3

现在，让我们随便选取一根线条。如果这根线条的两端是两个不同的顶点，那么把这根线条的长度缩短为 0，把两个端点合并为一个点。这样一来，顶点数和边数都减少了一个。如果这根线条的两端是同一个顶点，换句话说这根线条其实是一个“一边形”，或者说是一个“圈”，那么直接把这根线条删掉，边数和区域数都将减少一个。无论哪种情况，$V-E+F$ 的值都是不变的。如图 4 所示，把所有的线条都删掉后，整个平面上就只剩一个孤点和一个空荡荡的区域了，由此可知$V-E+F=2$。

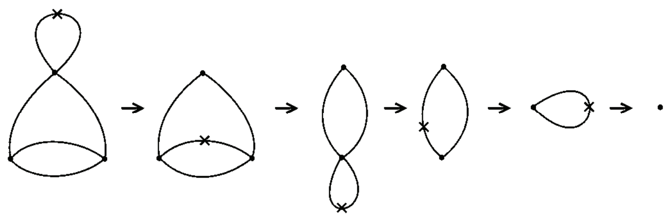


图 4

欧拉公式的实用价值非常高。在单面的印刷电路板中，线路是不允许相交的。假设电路板上有 A、B、C、D 四个点，有办法在每两个点之间都连接一条线路吗？图 5 的左边是一次失败的尝试，A、C 的连线和 B、D 的连线在中间相交了；不过稍作修正，



就能得到一个满足要求的布线方案。

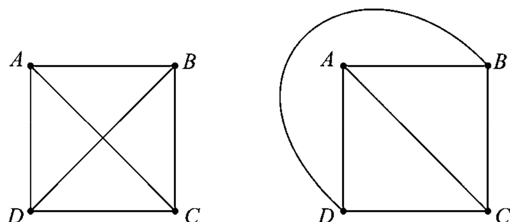


图　5

如果点的数目增加到5个,还能不交叉地连接所有的点对吗？欧拉公式告诉我们，这是绝对不可能办到的。首先注意到，在统计所有区域（包括最外边那个无穷大的区域）的边数之和时，每根线条都会被计算两次，因而所有区域的平均边数就是 $\frac{2E}{F}$ 。如果 5 个点之间两两连线，一共会产生 10 根线条。如果连线不相交，则它应该有 $E-V+2=10-5+2=7$ 个区域，于是每个区域平均拥有 $\frac{20}{7}$ 条边。这说明该图中至少存在一个边数小于 3 的区域。但是，我们只允许每两点之间连一条线，因此绝不会产生“一边形”或者“两边形”。因此，5 个点之间两两连线，线条将会不可避免地相交。

另外，在一个图形中，如果已知 V 、E 、F 三个数之中的任意两个，我们可以根据欧拉公式反过来求出第三个数。图 6 中，一个圆把平面分成了 2 份，两个圆把平面分成了 4 份，三个圆则把平面分成了 8 份。规律似乎很明显：n 个圆能够把平面分成 2^n 个区域。事实上真的如此吗？当 $n=4$ 时，例外发生了——此时整个平面只有 14 个区域。（回想我们在第 13 节讲的，当 $n=4$ 时，仅仅由圆构成的维恩图是不存在的。）

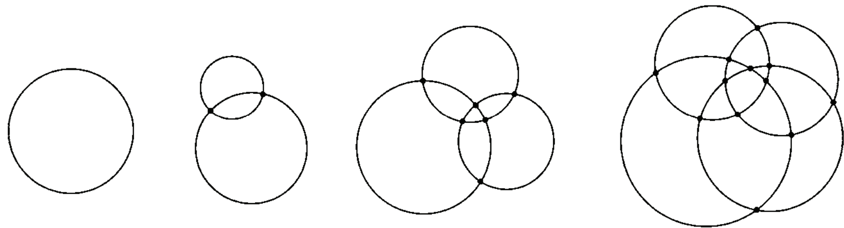


图　6



那么，这个数列的规律究竟是什么？ n 个圆两两相交，将会把整个平面分成多少块？利用欧拉公式，我们可以瞬间解决这个问题。由于每两个圆之间都有两个交点，因此顶点数 $V=n(n-1)$；由于每个圆被切成了 $2(n-1)$ 条弧，因此 $E=2n(n-1)$。于是，$F=E-V+2=2n(n-1)-n(n-1)+2=n^2-n+2$。巧的是，当 n 分别为 1、2、3 时，n^2-n+2 的值正好是 2、4、8，好一个误导人的数列！

要比结论的误导性，什么也比不过下面这个例子。圆上有 n 个点，两两之间连线后，最多可以把整个圆分成多少块？

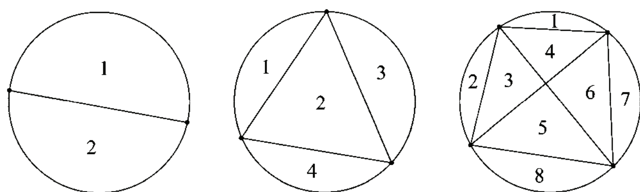


图 7

如果圆上只有一个点，圆内将不会产生任何线段，整个圆仍然是完整的一块。图 7 显示的就是 n 分别为 2、3、4 的情况。可以看到，圆分别被划分成了 2 块、4 块、8 块。规律似乎非常明显：圆周上每多一个点，划分出来的区域数就会翻一倍。事实上真的是这样吗？让我们看看当 $n=5$ 时的情况（见图 8）。

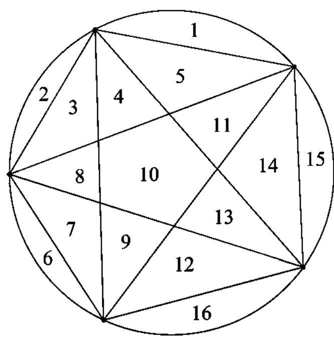


图 8

果然不出所料，整个圆被分成了 16 块，区域数依旧满足 2^{n-1} 的规律。此时，大家都会觉得证据已经充分，不必继续往下验证了吧。偏偏就在 $n=6$ 时，意外出现了（见图 9）。

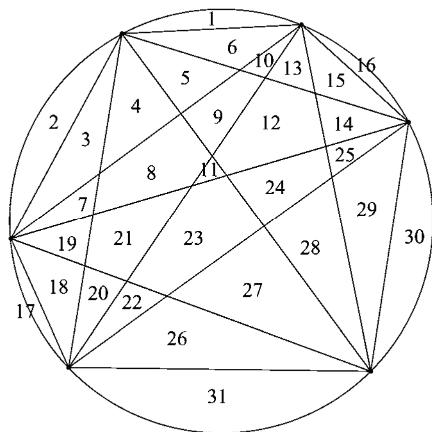


图　9

此时，区域数只有 31 个。那么，这个数列规律究竟是什么呢？这回，欧拉公式又帮上了大忙。圆周上每四个点交叉相连，就会在圆内产生一个交点，因此圆内一共有 C_n^4 个交点。加上圆周上本身的 n 个点，可得图中的总顶点数 $V=C_n^4+n$ 。圆内的每个顶点度数都为 4，圆周上的每个顶点度数都是 $n+1$ ，因此图中顶点的度数之和为 $4C_n^4+n(n+1)$ 。由于每根线条都贡献了两个度，因而图中的总线条数就是 $E=\frac{4C_n^4+n(n+1)}{2}=2C_n^4+\frac{n(n+1)}{2}$ 。因此，图中的总区域数就是：

$$\begin{aligned}F&=E-V+2\\&=\left(2C_n^4+\frac{n(n+1)}{2}\right)-\left(C_n^4+n\right)+2\\&=C_n^4+\frac{n(n-1)}{2}+2\end{aligned}$$

它可以写成 $C_n^4+C_n^2+2$ ，去掉圆外面那个无限大的区域，圆内部的区域数也就是 $C_n^4+C_n^2+1$ 。其实这个答案也不难理解：每画出一条新线段后，假如这条新线段与原来已有的线段产生了 k 个新交点，那么圆内就会新增加 $k+1$ 块区域。由于没有画任何线段时，圆内的区域数为 1，因此最终总的区域数就是 1 加上所有交点的个数，再加上所有线段的数量，也就是 $C_n^4+C_n^2+1$ 了。

关键在于， $C_n^4+C_n^2+1$ 又可以重写成 $C_{n-1}^4+C_{n-1}^3+C_{n-1}^2+C_{n-1}^1+1$ ，也就是杨辉三角（见图 10）第 n 行的前 5 个数之和。由于杨辉三角前 5 行都没超过 5 个数，因此当 n 是



小于等于 5 的正整数时，$C_{n-1}^4+C_{n-1}^3+C_{n-1}^2+C_{n-1}^1+1$就相当于是杨辉三角第 n 行所有数全部相加的结果。而杨辉三角第 n 行的所有数之和正好就是 2^{n-1}，于是便诞生了数学中最具误导性的“伪规律”。

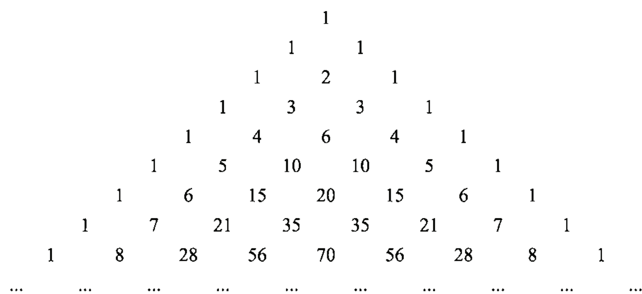


图 10

31. 定宽曲线与蒲丰投针实验

想象你在搬家时，需要让一个椭圆形的桌子通过一个笔直狭长的走廊。如果横着搬搬不过去，没准竖着搬就能搬过去了。毕竟，椭圆在不同方向上"宽度"是不一样的。不过，如果需要搬过去的是一个圆形的桌子，桌子的朝向就无所谓了。因为在各个方向上，圆的宽度都是一样的。

在数学中，违反直觉的东西太多了。你相信吗？除了圆以外，还有其他的平面几何图形，它在各个方向上的宽度也都一样！

让我们来构造一个满足要求的图形。如图 1，先画一个边长为 1 的等边三角形 *ABC*。然后，将每条线段两头各向外延长 1 个单位的长度，得到 *D*、*E*、*F*、*G*、*H*、*I* 这六个点。现在，以 *A* 为圆心，*AD* 为半径画弧，把 *D*、*E* 两点连接起来；再以 *B* 为圆心，*BE* 为半径画弧，把 *E*、*F* 两点连接起来；类似地，再分别以 *C*、*A*、*B*、*C* 为圆心，以 *CF*、*AG*、*BH*、*CI* 为半径画弧，把剩下的点连接起来。

这个似圆非圆的图形就满足，在各个方向上的宽度都是一致的。从图 2 中容易看出，在任意一个地方，这个图形的宽度都等于 3。

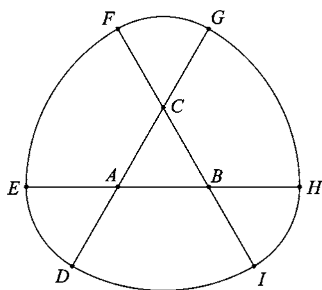


图　1

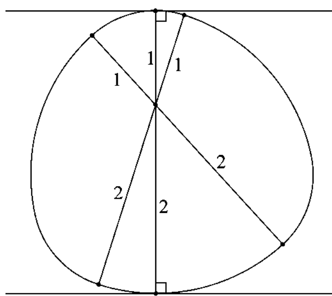


图　2



如果把若干个以此图形为横截面的铅笔垫在一块木板下，木板同样会毫无颠簸地向前滚动，和圆柱形滚轮的效果没什么两样。同样满足要求的图形还有很多很多，斯坦利·拉比诺维茨（Stanley Rabinowitz）甚至给出了一个非常复杂的八次曲线，它也满足宽度处处相同的性质。在数学中，我们把这种各处宽度都相同的平面几何图形叫做“定宽曲线”。

圆的很多性质都可以扩展到所有的定宽曲线中。例如，定宽曲线上任意两点之间的距离都不会超过图形的定宽。这又带来了下面这个有趣的“冷知识”：为了不让下水道井盖掉进下水道里，除了圆形井盖以外，所有定宽曲线形状的井盖也都是满足要求的。巴比尔（Barbier）定理则给出了定宽曲线的另一个漂亮的性质：如果一个定宽曲线的宽度为 d，那么它的周长就是 $\pi \cdot d$（正如圆的周长与直径的关系一样）。换句话说，取一些宽度相同但形状不同的定宽曲线，它们在地上滚动一周后，都将会前进相同的距离。

在第 14 节讲数学常数时，我们就已经谈到过 π 了。我们说过，圆周率 π 经常出现在一些和它毫无关系的场合中。其实，我们当时并没有提到最典型的一个例子——蒲丰投针实验。这是由 18 世纪法国著名数学家蒲丰（Comte de Buffon）在把微积分引入概率论时提出的：假设地板上画着一系列间距为 1 的平行线（见图 3），把一根长度为 1 的针扔到地上，则这根针与地板上的平行线相交的概率是多少？答案非常出人意料：这个概率为 $\frac{2}{\pi}$。

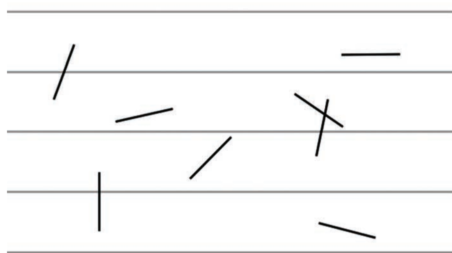


图 3

利用一些微积分知识，我们可以很快证明这一结论。假设一根针的中心与地板上最近的直线距离为 x，那么 x 的取值一定在 0 到 $\frac{1}{2}$ 之间。此时，只要图 4 中所示的夹角 θ 不超过 $\arccos\left(\frac{x}{1/2}\right)$，这根针就会和直线相交。

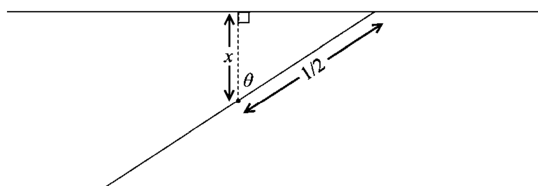


图　4

我们建立一个平面直角坐标系，其中横轴代表针的中心与最近直线的距离，取值范围是从 0 到 $\frac{1}{2}$；纵轴则代表针的倾斜角度，取值范围是从 0 到 $\frac{\pi}{2}$。如果把针在地板上的分布情况用这么一个矩形区域中的点集来表示，那么针会与平行线相交的情况就是图 5 中的阴影部分。

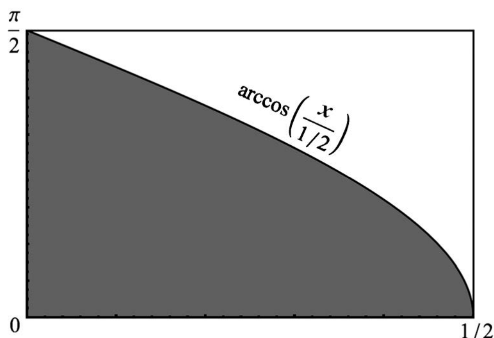


图　5

利用微积分不难算出，阴影部分的面积是 $\int_0^{1/2} \arccos\left(\frac{x}{1/2}\right) \mathrm{d}x = \frac{1}{2}$。而整个矩形的面积是 $\frac{\pi}{2} \times \frac{1}{2} = \frac{\pi}{4}$，前者占后者的 $\frac{2}{\pi}$。这就说明，针与平行线相交的概率是 $\frac{2}{\pi}$。

不过，即使看到了结论的证明过程，大家或许还是感到很不理解：结论里的 π 究竟是哪里来的呢？这个结论是否有一个更加直观的解释呢？

现在，让我们来考虑任意长度甚至是任意形状的针，或者叫铁丝更为恰当一些。如图 6，把这样的铁丝扔在地板上，铁丝与平行线有可能相交不止一次。我们有这样一个神奇的结论：给定一根弯弯曲曲的铁丝，把它扔在地板上后，它与平行线的平均交点数量只与铁丝本身的长度有关。铁丝越长，平均交点数也会越大，两者成一个正



比关系。下面有一个直观的证明思路。我们可以把这根铁丝看作是很多条短小的直线段组成的。那么，在大量的投铁丝实验，比如 1 亿次实验后，铁丝与平行线相交的总次数，就等于所有的小线段在所有 1 亿次实验中与平行线相交次数的总和。但是，每一条小线段的形状都是相同的，并且大量实验后，它们的落点最终都会均匀地分布在整个地板上（即使这些小线段之间是首尾相连的）。因此，在这 1 亿次实验中，每条小线段各自与平行线的相交总次数都是大体相同的。铁丝越长，铁丝所含的小线段越多，铁丝与平行线的总交点数也就会越多。自然，平均每次实验中铁丝与平行线的交点数，也就与铁丝的长度成正比了。也就是说，假设铁丝的长度是 L ，则铁丝与平行线的平均交点个数就是 $c \cdot L$ ，其中系数 c 是一个常数。

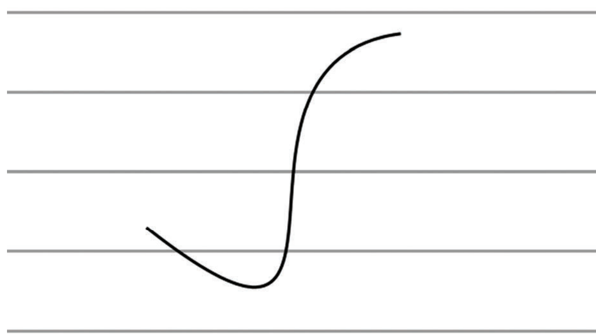


图 6

这个常数是多少呢？为了求出这个常数，我们只需要考虑一些特殊的情况。注意到，把一根长度为 π 的铁丝弯成一个直径为 1 的圆，再把它扔到地上之后，它与这组平行线总有两个交点。这就是说， $c \cdot \pi = 2$ ，即 c 等于 $\frac{2}{\pi}$ 。那么，一根单位长度的针与平行线的交点个数的期望值就是 $\frac{2}{\pi}$ ；而由于这根针与平行线不可能有两个或两个以上的交点，因此这个数值就相当于是针与平行线相交的概率了。

好了，真正神奇的地方来了。由于“直径”为 1 的定宽曲线与平行线也总有两个交点，因此它的周长必然也是 π 。我们就这样证明了巴比尔定理，而巴比尔定理本来和概率论没有半点关系！



32. 来自不同领域的证明

说到数学证明，不得不谈到数学界的一篇奇文。1987 年，斯坦・瓦根（Stan Wagon）在《美国数学月刊》上发表了一篇题目为《一个矩形剖分定理的十四种证法》的论文，论文中提到了这么一个定理：

如果一个矩形可以分割为若干个小矩形，每个小矩形都有至少一边为整数长，则原矩形同样有至少一边为整数长。

换句话说，用至少有一边的长度是整数的小矩形拼成一个大矩形，大矩形也一定有至少一条整数长的边。这个命题看似简单，想到证明方法却并不容易。斯坦・瓦根竟然给出了十四种完全不同的证明方法，每一种证明方法都非常巧妙。我选择了其中六个最具代表性的证明，和大家一同分享。

我们所要介绍的第一个证明是我觉得最巧妙的证明方法。证明的关键在于下面这个引理：像国际象棋棋盘一样对整个平面黑白染色，那么与两坐标轴平行放置且至少一边长为偶数个单位的矩形一定覆盖了相同面积的黑色区域和白色区域。原因很简单，如图 1，无妨假设这个矩形的水平方向上的边是偶数个单位长，那么该矩形中的每个横条显然都覆盖了相同面积的黑白两色区域。对于竖直边的边长为偶数的矩形，也是同样的道理。

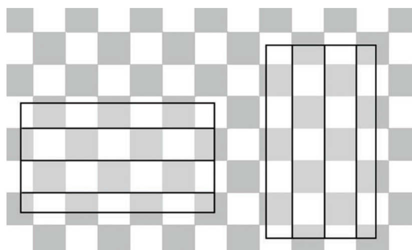


图　1



下面，我们把平面分成$\frac{1}{2}\times\frac{1}{2}$大小的正方形，并且像上面那样对其进行黑白二染色，然后像图 2 那样，将整个大矩形对齐坐标轴放在平面上，左下角和原点重合。这个矩形内的每个小矩形都有至少一条整数边，也即至少有一边的长度是$\frac{1}{2}$的偶数倍，因此每个小矩形都覆盖了相同面积的黑色区域和白色区域。这样，整个大矩形也就覆盖了相同面积的黑色区域和白色区域。

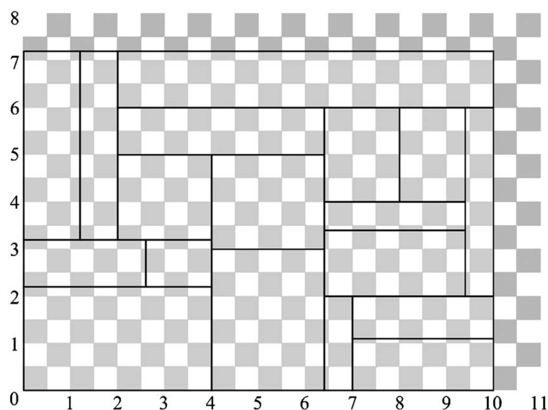


图 2

下面我们将说明，对于一个左下角与原点对齐的矩形，如果右上角的顶点 (x, y) 的两个坐标值都不是整数，那么这个矩形所覆盖的黑色区域的面积一定大于白色区域。如图 3，除去黑白相等的“整”的部分，最后剩下的就是最右上角的那个横纵两个方向均未被填满的1×1正方形有待讨论。

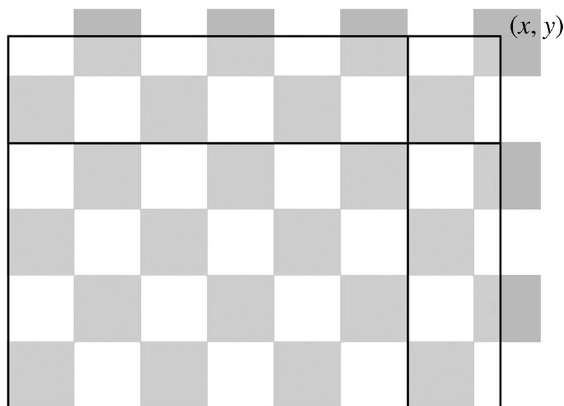


图 3



如图 4，坐标 (x, y) 的位置只有 A、B、C 三种可能。如果坐标 (x, y) 位于区域 A，很容易看出黑色面积比白色面积大；如果坐标 (x, y) 位于区域 B，也很容易看出黑色面积比白色面积大；如果坐标 (x, y) 位于区域 C，则矩形没有覆盖到的拐角形区域中白色面积更大，因而矩形内的黑色面积就更大一些。

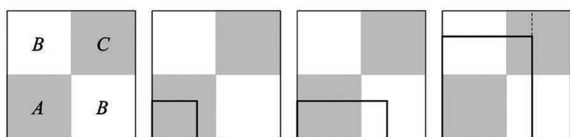


图 4

如果要让整个矩形覆盖相同面积的黑色区域和白色区域，x 和 y 至少有一个是整数才行，而这就是我们所要证明的结论了。

利用数学归纳法，我们可以得到另一个思路完全不同的初等的证明。如图 5，我们首先把每个小矩形都分割成单位宽度的长条。这样的话，大矩形里就只有两种小矩形：宽为 1 的竖条状矩形（图 5 中的浅色矩形）和高为 1 的横条状矩形（图 5 中的深色矩形）。我们对浅色矩形的个数施加归纳。随便选择一个浅色的矩形（例如图 5(2) 中的阴影矩形），增加它的高度，让它“穿过”它头顶上的深色矩形（把它正上方的深色矩形截断），直到这根竖条状矩形的顶端碰到了另一个浅色矩形的底端。把后者作为新的操作对象，继续增加其高度，必要时再次更换操作对象，直到达到整个大矩形的上边界。我们用同样的方法让最初所选的阴影矩形向下“生长”到大矩形的下边界。

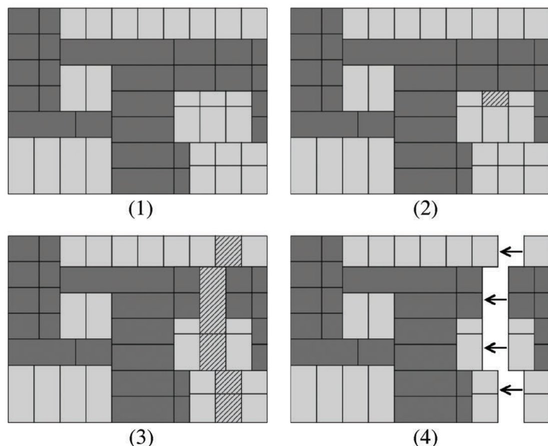


图 5



注意到在此过程中，浅色矩形始终保持着单位宽度，深色矩形始终保持着单位高度。整个过程结束后，深色矩形的个数变多了，但浅色矩形的个数不变。此时我们得到了一条上下贯穿整个大矩形的浅色矩形链。把它们擦掉，将右半部分左移一个单位，重新拼成一个大矩形。新的大矩形高度不变，宽度减 1，因而原来的整数边还是整数，非整数边仍然不是整数。同时，浅色矩形的个数减少了。反复进行这样的操作，总有一个时候大矩形里只剩下深色的矩形（则原大矩形高度显然为整数），或者某次操作后所有矩形都被去掉了（则原大矩形宽度为整数）。

借用这种方法我们还可以得到一个颇有喜剧效果的反证法。假设大矩形的长度和宽度都不是整数，那么每一步操作后，它们仍然是非整数，这表明大矩形里不可能只剩一种颜色的小矩形，于是我们可以无限制地调用上面的操作。最后的结果是：我们得到了一个用整数长或整数宽的小矩形拼成的一个大矩形，而这个大矩形的横边竖边都小于 1！这显然是荒谬的。

接下来，我将要给大家介绍第三种证明方法——图论方法。如图 6，首先，找出图中所有的“交叉点”（包括整个大矩形的四个顶点），然后按照下面的方式把它们连起来：对于每个小矩形，如果它的两条水平边的长度是整数，就用上下两条线分别连接水平方向上的两对点；如果它的两条竖直边的长度是整数，就用左右两条线分别连接竖直方向上的两对点；如果它的四条边的长度都是整数，只需连接其中一组对边即可。这样的话，每个矩形都会产生两条连线，矩形的四个顶点各被用过一次。于是，我们得到了一个由若干顶点和这些顶点之间的连线构成的图。在这个图中，大矩形的四个顶点的度数为 1；由于其他每个交叉点都同属于两个小矩形或者四个小矩形，因此其余顶点的度数都是 2 或者 4。下面，我们把这个矩形放在平面直角坐标系中，大矩形的左下角对齐原点 (0,0) 。现在，从原点出发，沿着我们所画的线条行走，并且把沿途走过的线条都擦掉。显然，我们走到的每个点的两个坐标均为整数。

我们是从一个度数为 1 的顶点出发的，显然不可能再回到出发点了，只能沿着图中的线条漫无目的地游荡。但是，图中的总边数是有限的，总有一个时候我们将会无路可走。注意到，我们绝不可能在度数为 2 或者 4 的地方无路可走，因为度数为偶数也就意味着这个顶点“有进必有出”。因此，这趟旅程的终点必然会落在另一个度数为 1 的点上。这个终点一定是大矩形的另一个角，因而它的两个坐标值均为整数。于是命题得证。

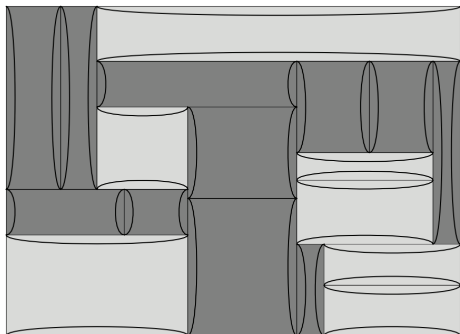


图 6

利用类似的思路，安德烈·内普（Andrei Gnepp）曾给出一个更简单的证明：由于每个小矩形都有至少一对整数长的边，因而一个小矩形的四个顶点中，两个坐标值均为整数的顶点只可能有 0 个、2 个或者 4 个。把它们全部加起来，符合条件的总顶点数 S 仍然是偶数。但是，这 S 个顶点中有些点是重复算过的。除了大矩形四个角上的顶点外，其他每个顶点都同属于两个矩形或者四个矩形，如果某个顶点的横纵坐标都是整数，则它将会被重复计算两次或者四次。假如我们有 S_1 个两坐标均为整数的顶点只被算过一次，有 S_2 个这样的点被算过两次，有 S_4 个这样的点被加了四次，则有 $S = S_1 + 2\times S_2 + 4\times S_4$。我们立即得出，$S_1$ 也是偶数。但我们已经有一个只被算过一次的点（即最左下角的点 $(0,0)$），那么 S_1 至少为 2，即至少还有一个两坐标均为整数并且只被算过一次的点，它即是大矩形的另一个角。

彼得·温克勒（Peter Winkler）也曾给出图论证明法的另一个变形。如图 7，还是把整个大矩形放在平面直角坐标系中，左下角和原点重合。现在，对于每个小矩形，如果它的两条水平边的长度是整数，就把这个矩形染成浅色，不过在上下两条边各留下一个很窄很窄的深色横条；如果它的两条竖直边的长度是整数，就把它染成深色，不过左右两条边各留下一个很窄很窄的浅色竖条；如果它的四条边的长度都是整数，就随便采用一种染色方案。那么，整个大矩形中要么存在一条从左边界到右边界的浅色路径，要么浅色区域没能连通左右边界，从而整个图中存在一条从下边界到上边界的深色路径。在前一种情况中，容易看出，这条路径穿过的所有竖直线段的 x 坐标都是整数，这就表明整个大矩形的宽度是整数；类似地，后一种情况也就表明，整个大矩形的高度是整数。



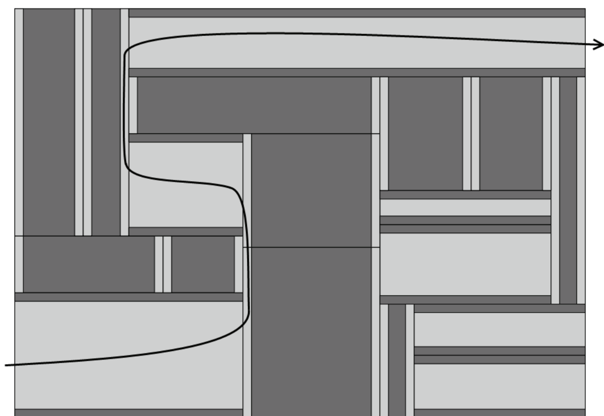


图 7

接下来是第四种证明方法，这种证明方法可就有些另类了。同样将矩形放置在平面直角坐标系中，左下角对齐原点 $(0,0)$ 。选取一个充分小的变量 t 。对于每个小矩形，把所有 x 坐标不是整数的竖直边都向右平移 t ,把所有 y 坐标不是整数的水平边都向上平移 t 。如果一个小矩形的竖直边的长度是整数，那么它的两条水平边要么都被平移了 t ，要么都没有被平移；如果一个小矩形的水平边的长度是整数，那么它的两条竖直边要么都被平移了 t ，要么都没有被平移。总之，一个宽和高分别为 w 、 h 的小矩形，变化之后的新的面积只可能是 $w\times(h\pm t)$ 、 $(w\pm t)\times h$ 、 $w\times h$ 中的一种，它一定是一个关于 t 的一次函数（包括常函数的情况）。因此，整个大矩形的新面积也是一个关于 t 的一次函数。然而，如果大矩形的两条边的长度 a 、 b 都不是整数，大矩形的面积将会变为 $(a+t)(b+t)$ ，这是一个关于 t 的二次函数。因此，大矩形至少会有一条整数长的边。

下面轮到真正怪异的证明方法登场了。借助看上去与这个问题毫不相干的微积分知识，我们能够迅速证明这一结论。还是把矩形放在平面直角坐标系中，左下角对齐原点 $(0,0)$ 。考虑函数 $\mathrm{e}^{2\pi\mathrm{i}(x+y)}$ 在每个小矩形上的积分：

$$
\begin{aligned}
&\int_{x_0}^{x_1}\int_{y_0}^{y_1}\mathrm{e}^{2\pi\mathrm{i}(x+y)}\mathrm{d}y\mathrm{d}x \\
=&\int_{x_0}^{x_1}\mathrm{e}^{2\pi\mathrm{i}x}\mathrm{d}x\cdot\int_{y_0}^{y_1}\mathrm{e}^{2\pi\mathrm{i}y}\mathrm{d}y \\
=&\frac{1}{(2\pi\mathrm{i})^2}\left(\mathrm{e}^{2\pi\mathrm{i}x_1}-\mathrm{e}^{2\pi\mathrm{i}x_0}\right)\left(\mathrm{e}^{2\pi\mathrm{i}y_1}-\mathrm{e}^{2\pi\mathrm{i}y_0}\right)
\end{aligned}
$$

显然，这个式子等于 0 当且仅当 $x_1 - x_0$ 和 $y_1 - y_0$ 中至少一个是整数，也即该小矩形至少有一边的长度是整数。考虑函数在整个大矩形上的积分，它可以拆成各个小矩形上的积分的和，因此结果仍然是 0。这说明，大矩形中也有至少一条整数长的边。

令人更加难以置信的是，这个结论竟然还有一种数论证法。证明的关键就在于，质数有无穷多个。给定一个满足要求的大矩形，如果你宣称它的四条边的长度都不是整数，它们都多出了大小至少为 ε 的“零头”。那么，我总能找出一个足够大的质数 p，使得 $\frac{1}{p} < \varepsilon$。然后我们将说明，大矩形其中有一条边的长度除去整数部分后的“零头”不会超过 $\frac{1}{p}$，由此产生矛盾。这样的话，至少有一边恰好是整数长才行。

仍然是把大矩形放在平面直角坐标系中，左下角对齐原点 $(0,0)$。考虑所有形如 $\left(\frac{i}{p}, \frac{j}{p}\right)$ 的点所形成的点阵（其中 i、j 均为整数）。我们需要把整个点阵平移到一个合适的位置，使得点阵中没有点恰好落在小矩形的边界上。这总是可以办到的，例如，我们算出每个小矩形的每条横边到点阵中离它最近的点的距离，取所有这些最近距离中最小的非 0 的值，然后在竖直方向上将点阵移动一个比这还小的距离；另一个方向亦是如此。

假如数轴上分布着间隔为 $\frac{1}{p}$ 的点，容易看出，任取一个长度为整数的区间，只要这些点不与区间的端点重合，那么区间内所含点数一定是 p 的倍数。注意到每个小矩形内部所含的点数都是两个数的乘积，由于其中至少有一个数恰好是 p 的倍数，因此每个小矩形内都有 p 的倍数个点。那么，整个大矩形所含的点的个数（即所有小矩形所含点数之和）也是 p 的倍数。大矩形内的所有点的个数也是两个数的乘积，然而 p 是质数，因此两个数中至少一个含有因数 p。那么，对应的那条边也包含了 p 的倍数个点。这说明，这条边应该是整数长，最多有 $\frac{1}{p}$ 的误差。



33. 平分面积的直线

零点定理是大家平时生活中用惯了以至于反而觉得很陌生的一个定理。若函数 $f(x)$ 在区间 $[a,b]$ 连续，并且 $f(a)$ 与 $f(b)$ 一正一负，那在 (a,b) 之间一定存在某个 x，使得 $f(x)=0$。如果你从海拔为 -100 米的地方走到海拔为 400 米的地方，那不管你是怎么走的，都一定会有某一时刻恰好位于海平面高度。另一个比较隐蔽一些的应用便是，对任意一个凸多边形，总存在一条直线把它分成面积相等的两份。考虑一条竖直直线从左至右扫过整个凸多边形，则凸多边形位于直线左边的那部分面积由 0 逐渐增大为整个凸多边形的面积，直线右侧的面积则由最初的整个凸多边形面积渐渐变为 0。若把直线左侧的面积记为 $f(x)$，直线右侧的面积记为 $g(x)$，则随着直线位置 x 的变化，$f(x)-g(x)$ 的值由一个负数连续地变为了一个正数，它一定经过了一个零点。这表明，在某一时刻一定有 $f(x)=g(x)$。

大家或许曾经想过这样一个问题：对于任意一个凸多边形，我们总能用两条互相垂直的直线把它的面积分成四等份吗？答案是肯定的。如图 1，利用前面的结论，我们能找到一条直线 l_1，它把整个凸多边形分成上下相等的两份；类似地，我们能找到唯一的一条与 l_1 垂直的直线 l_2，使得它恰好把整个凸多边形分成左右相等的两份。注意，现在我们有 $A_1+A_2=A_2+A_3=A_3+A_4=A_4+A_1$，由此还可以立即知道 $A_1=A_3$ 并且 $A_2=A_4$，但这都还不足以保证四块面积全都相等。怎么办呢？注意，我们前面假定直线 l_1 是一条水平直线。事实上，l_1 每取一个方向，我们都能用上面的方法得到一个具有相同性质的新构造。现在，我们将直线 l_1 的方向顺时针旋转 90 度。考虑整个过程中 A_1-A_2 的值的变化：旋转后的 A_1-A_2 恰好就是旋转前的 A_2-A_3，而 A_1 和 A_3 又是相等的。于是我们发现，旋转前后 A_1-A_2 的值恰好互为相反数！这表明，在直线 l_1 旋转的过程中，一定有一瞬间满足 $A_1-A_2=0$，这一刻的 l_1 和 l_2 便是两条互相垂直并把图

形四等分的直线。为了证明这个结论，我们三次嵌套地使用了零点定理！

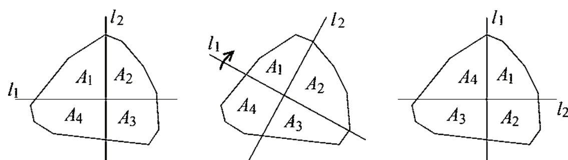


图 1

故事并未到此结束。我们还有这样的定理：对于任意一个凸多边形，总能用三条交于一点的直线把它的面积分成六等份。

如图 2，先用直线 l_1 把图形分成上下相等的两半。对于 l_1 上的任意一点 P，总存在唯一一组的四条射线，它们和直线 l_1 一起恰好把图形分成六等份。我们把 r_1 和 l_1 的夹角记做 α，把 r_3 和 l_1 的夹角记做 β。现在，考虑点 P 从 l_1 最左边向最右边移动，则角 α 由 180 度慢慢变成 0 度，角 β 则从 0 度慢慢变成 180 度，因此在此过程中必然有 $\alpha=\beta$ 的时刻。此时 r_1 和 r_3 就在一条直线上了。接下来，将 l_1 的方向顺时针旋转 180 度，同时不断调整点 P 的位置，保持 r_1 和 r_3 始终在一条直线上。最终得到的构造将会和刚才一样，只不过 r_2 和 r_4 交换位置了：原来 r_4 在 r_2 延长线的顺时针方向，现在 r_4 跑到了 r_2 的延长线的逆时针方向，前后两个有向角的角度互为相反数。因此，在 l_1 旋转的过程中，必然有某个时刻 r_2 的延长线和 r_4 正好重合。此时，l_1、r_1 和 r_3 所在的直线、r_2 和 r_4 所在的直线就是把凸多边形面积分成六等份的 3 条直线。

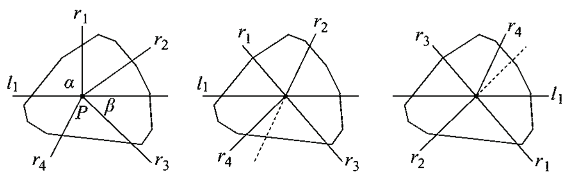


图 2

这个证明过程真可谓是把零点定理用到了炉火纯青的地步。大家不妨数一数，在这个证明过程中，我们一共嵌套地用到了多少次零点定理！



34. 小合集（二）：图形证明

你所见过的最短的证明有多长？一句话？一个字？事实上，真正最直观、最简单的证明过程连一个字都不用，这便是证明的最高境界——“无字证明”。在证明一个数学命题，尤其是一些与数列相关的命题时，比起大段大段的文字来，一张不附带任何文字的图片往往更能说明问题。

先来看一个有趣的智力题吧：如何把图 1 中的图形分成大小形状完全相同的四等份？

如果你是第一次听说这个题目，你一定会觉得答案异常巧妙（见图 2）。

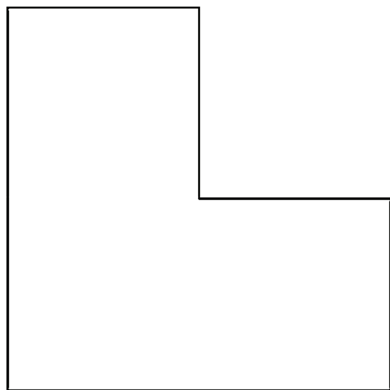


图 1

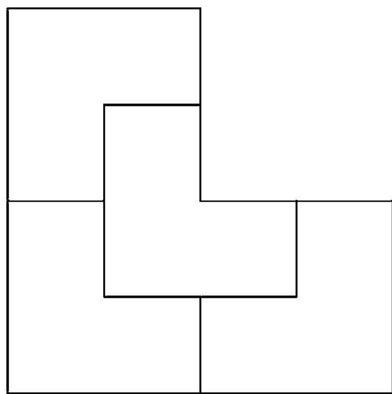
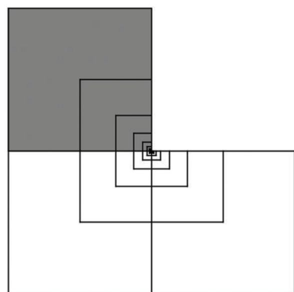


图 2

有趣的是，我们分割出来的每一小份正好都与原来的整个图形是相似的，因此我们可以选取其中一小份再次四等分，并无限地这样做下去，从而不断得到图形总面积的 $\frac{1}{4}$、$\frac{1}{16}$、$\frac{1}{64}$ 等等。我们立即得到图 3。



$$\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\cdots=\frac{1}{3}$$

图 3

于是，$\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\cdots=\frac{1}{3}$这一结论变得几乎是显而易见的了！

除了“拐角形”以外，很多其他的几何图形也具有这样的性质：它包含了四个更小的自己。图 4 所示的是两个梯形，它们的下底都是上底的两倍，左边那个梯形的两底角分别是 45 度和 90 度，右边那个梯形的两个底角都是 60 度。这两个图形也可以分割成和原图形相似的四份。

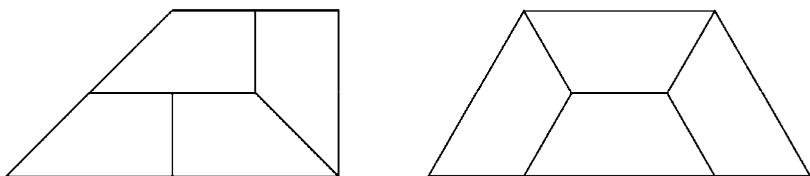
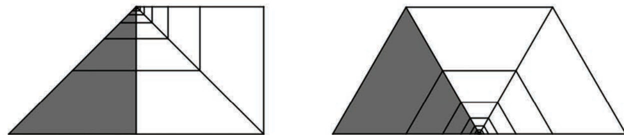


图 4

它们则对应了$\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\cdots=\frac{1}{3}$的如图 5 所示的另外两种图形证明。



$$\frac{1}{4}+\frac{1}{16}+\frac{1}{64}+\cdots=\frac{1}{3}$$

图 5



不过，这都是以$\frac{1}{4}$为底的几何级数。我们自然会想，有没有什么图形能够用来证明别的等式，比如$\frac{1}{3}+\frac{1}{9}+\frac{1}{27}+\cdots=\frac{1}{2}$呢？按照上面的思路，我们首先需要找到这样的图形，它可以分成三个和自身相似的小块。这样的图形并不是没有，如图 6，有 30 度角的直角三角形以及长宽比为$\sqrt{3}:1$的矩形都满足要求；但可惜，图中各个小块的排列位置很不理想，不断地选取整个图形的$\frac{1}{3}$、$\frac{1}{9}$、$\frac{1}{27}$等等，并不能构成一个明显等于$\frac{1}{2}$的区域。

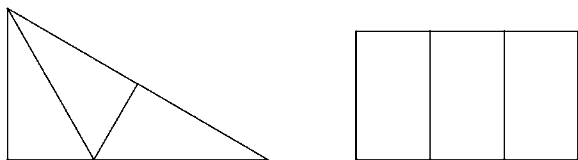
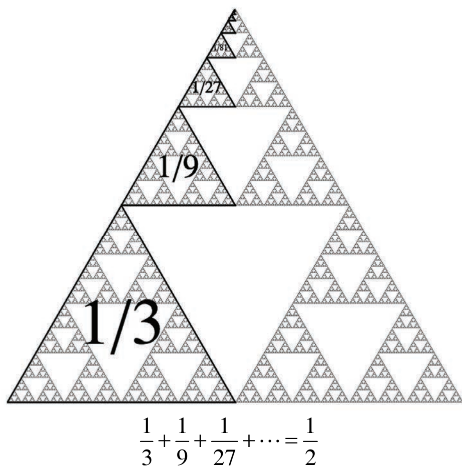


图　6

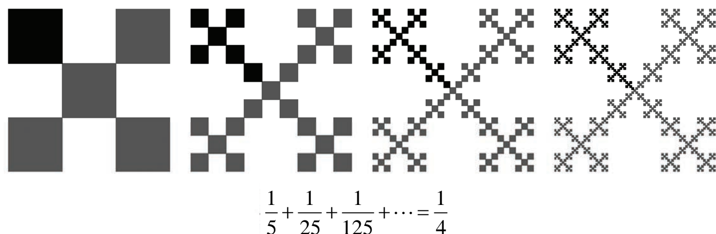
怎么办呢？别忘了，我们还有一种非常特殊的“自相似”图形——分形图形。在第 16 节讲分形图形时，我们特别提到过，谢尔宾斯基三角形是最经典的分形图形之一，在后面将会有意想不到的用处。现在是时候兑现了。如图 7 所示，利用谢尔宾斯基三角形，我们能够立即说明$\frac{1}{3}+\frac{1}{9}+\frac{1}{27}+\cdots=\frac{1}{2}$。



$$\frac{1}{3}+\frac{1}{9}+\frac{1}{27}+\cdots=\frac{1}{2}$$

图　7

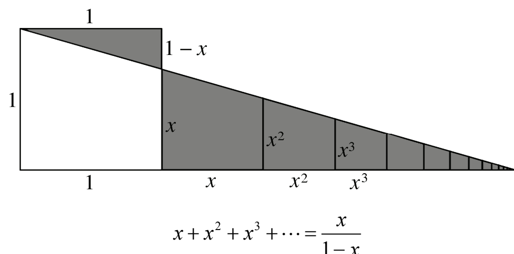
而一个叫做维则克（Vicsek）雪花的分形图形（见图 8）则可以用来证明 $\frac{1}{5}+\frac{1}{25}+\frac{1}{125}+\cdots=\frac{1}{4}$。



$$\frac{1}{5}+\frac{1}{25}+\frac{1}{125}+\cdots=\frac{1}{4}$$

图　8

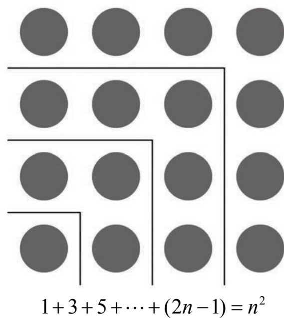
其实，几何级数公式还有另一种思路完全不同的图形证明方法，它对任何底数都是适用的。对于任意一个 0 到 1 之间的实数 x，我们都可以用一连串相似的梯形来构造 x^n。只需要注意到图 9 中两个阴影三角形相似，便有 $x+x^2+x^3+\cdots=\frac{x}{1-x}$。



$$x+x^2+x^3+\cdots=\frac{x}{1-x}$$

图　9

其他一些与数列求和有关的命题，也能用图形证明瞬间秒杀。图 10 所示的就是 $1+3+5+\cdots+(2n-1)=n^2$ 的图形证明。

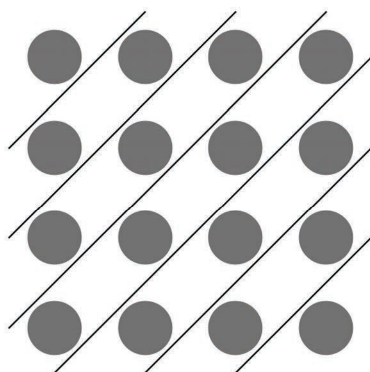


$$1+3+5+\cdots+(2n-1)=n^2$$

图　10



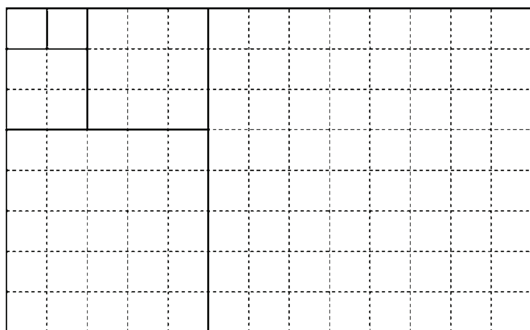
另一个有趣的例子则是 $1+2+\cdots+(n-1)+n+(n-1)+\cdots+1=n^2$，如图 11 所示。



$1+2+\cdots+(n-1)+n+(n-1)+\cdots+1=n^2$

图 11

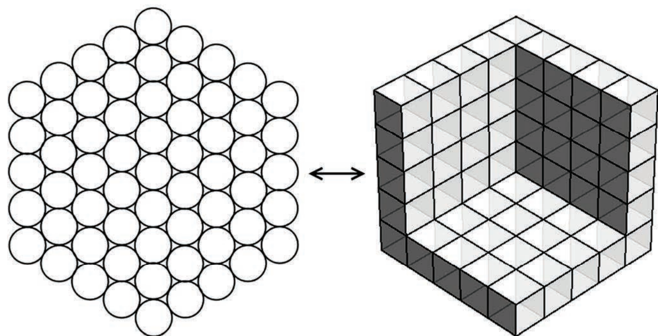
而斐波那契数列也有一个惊人的性质，即 $F_1^2+F_2^2+F_3^2+\cdots+F_n^2=F_n\cdot F_{n+1}$，从图 12 来看这几乎是显然的。



$F_1^2+F_2^2+F_3^2+\cdots+F_n^2=F_n\cdot F_{n+1}$

图 12

你知道吗？把可乐罐摆成边长为 n 的六边形阵，需要 $n^3-(n-1)^3$ 听可乐。用数学语言来说，就是第 n 个六边形数 h_n 等于 $n^3-(n-1)^3$。它的证明方法是我见过的最诡异的图形证明，见图 13。



$h_n = n^3 - (n-1)^3$

图　13

事实上，不仅仅是数列问题，一些更纯粹的代数问题也能转化为图形证明。我最喜欢的是下面这个问题：若 a 、b 、c 、d 都大于 0，求证 $\sqrt{a^2+b^2}+\sqrt{c^2+d^2} \geqslant \sqrt{(a+c)^2+(b+d)^2}$ 。我常常拿这个问题去考我的朋友们，搞得他们抓耳挠腮，怎么也证不出来。公布答案后，对方总会大叫上当，随即惊叹证明竟然如此简单，见图 14。

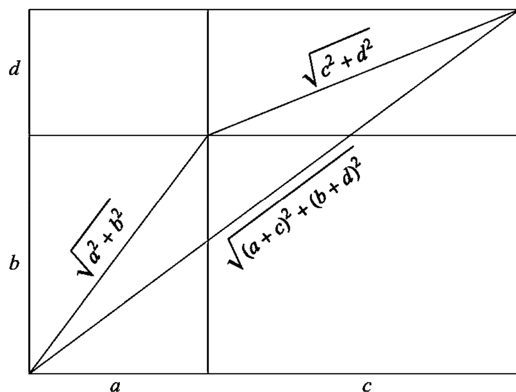


图　14

另一个有趣的数学事实是，如果两个分数的分子分母都是正数，则把它们的分子加在一起，分母加在一起，得到的新分数的大小一定在原来两个分数之间。它也能用图形方法迅速得证，而证明所用的图形竟然跟前一个问题中的图形一模一样。我们只需要把关注的重心从这几条斜线段的长度转移到它们的斜率即可，如图 15 所示。



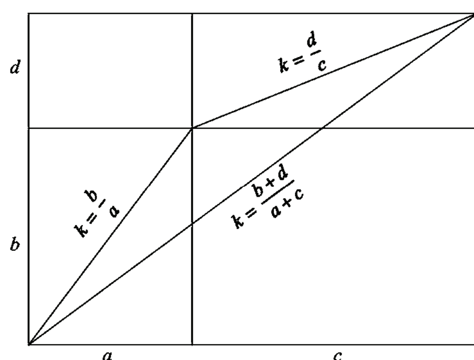
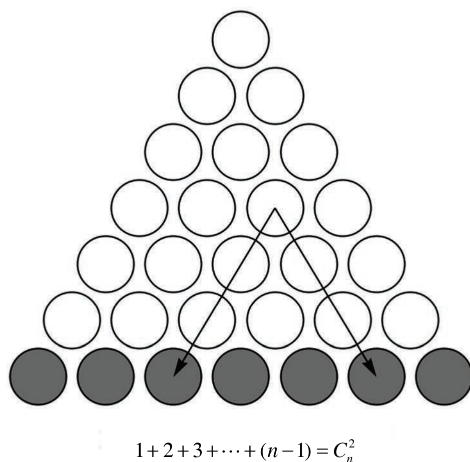


图　15

数学问答网站 MathOverflow 上曾经有人提问征集最漂亮的图形证明，得票最多的则是图 16 所示的这个证明。



$$1+2+3+\cdots+(n-1)=C_n^2$$

图　16

35. 生成函数的妙用

在这一节中，你将会看到，我们是如何用全新的数学工具去解决一个数学难题的。

有这么一个经典的概率问题：平均需要抛掷多少次硬币，才会首次出现连续两个正面？答案是 6 次。它的计算方法大致如下。

首先，让我们来考虑这样一个问题：k 枚硬币摆成一排，其中每一枚硬币都可正可反；如果里面没有相邻的正面，则一共有多少种可能的情况？这可以用递推的思想来解决。不妨用 $f(k)$ 来表示摆放 k 枚硬币的方案数。我们可以把这些方案分成两类：最后一枚硬币是反面，或者最后一枚硬币是正面。如果是前一种情形，则我们只需要看前 $k-1$ 枚硬币有多少摆法就可以了；如果是后一种情形，那么倒数第二枚硬币必须是反面，因而这种情形下的方案数就取决于前 $k-2$ 枚硬币的摆放方案数。因此我们得到，$f(k)=f(k-1)+f(k-2)$。由于摆放一枚硬币有两种方案，摆放两枚硬币有三种方案，因而事实上 $f(k)$ 就等于 F_{k+2}，其中 F_i 表示斐波那契数列 1, 1, 2, 3, 5, 8, …的第 i 项。

而“抛掷第 k 次才出现连续两个正面”的意思就是，最后三枚硬币是反、正、正，并且前面 $k-3$ 枚硬币中正面都不相邻。因此，在所有 2^k 种可能的硬币正反序列中，只有 F_{k-1} 个是满足要求的。也就是说，我们有 $\frac{F_1}{4}$ 的概率在第二次抛币就得到了连续两个正面，有 $\frac{F_2}{8}$ 的概率在第三次得到连续两个正面，有 $\frac{F_3}{16}$ 的概率在第四次得到连续两个正面……因此，我们要求的期望值就等于：

$$\sum_{k=1}^{\infty} k\times\frac{F_{k-1}}{2^k}=2\times\frac{1}{4}+3\times\frac{1}{8}+4\times\frac{2}{16}+5\times\frac{3}{32}+6\times\frac{5}{64}+7\times\frac{8}{128}+\cdots$$

不过，怎样求解这个无穷级数的和呢？你会发现，数学归纳法、求通项公式、错位相消等传统的数列求和方法此时似乎都没有了用武之地。现在，让我们来介绍一种更加强大的数列处理工具——生成函数。



让我们先来说说什么是生成函数吧。生成函数就是对数列进行编码的一种方式。我们可以用一个无穷级数 $a_1 \cdot x^1 + a_2 \cdot x^2 + a_3 \cdot x^3 + \cdots$ 把整个数列的全部信息装进去，其中第 i 次项系数就表示数列的第 i 项。因此，斐波那契数列的生成函数就可以写成：

$$g(x) = x + x^2 + 2x^3 + 3x^4 + 5x^5 + 8x^6 + 13x^7 + \cdots$$

厉害就厉害在，我们可以把生成函数表示成一个更简单的形式。先来看看 $g(x) \cdot x$ 的结果：

$$g(x) \cdot x = x^2 + x^3 + 2x^4 + 3x^5 + 5x^6 + 8x^7 + 13x^8 + \cdots$$

再看看 $g(x) + g(x) \cdot x$ 的结果：

$$g(x) + g(x) \cdot x = x + 2x^2 + 3x^3 + 5x^4 + 8x^5 + 13x^6 + 21x^7 + \cdots$$

你会发现，斐波那契数列的递推性质，使得上面这行式子与 $g(x)$ 本身非常相像。事实上，如果把 $g(x)$ 的每一项都除以 x，再减去最前面多出来的 1，就能得到上面的这行式子了。因此，我们有：

$$g(x) + g(x) \cdot x = \frac{g(x)}{x} - 1$$

我们甚至可以就此解出 $g(x)$ 来：

$$g(x) = \frac{x}{1 - x - x^2}$$

于是，整个无穷级数 $g(x)$ 被我们化简为了一个关于 x 的代数式！注意，虽然这个等式只在 x 充分小（小到级数 $g(x)$ 收敛）的时候才有意义，不过这并不妨碍我们用这个代数式来代表斐波那契数列的生成函数。我们可以把斐波那契数列看做是生成函数的一个“展开”：

$$\frac{x}{1 - x - x^2} = x + x^2 + 2x^3 + 3x^4 + 5x^5 + 8x^6 + 13x^7 + \cdots$$

也就是说，这么一个小小的代数式就容纳了斐波那契数列的全部信息！

生成函数是如此地具有代表性，以至于在研究数列时，我们常常会给出它的生成函数。在网络在线数列百科全书 oeis.org 中，生成函数几乎是必不可少的一项。例如，在斐波那契数列的描述中，FORMULA 一栏的第一行就是 G.f.:x/(1-x-x^2)，说的就是斐波那契数列的生成函数。

更绝的是，我们还可以直接对数列的生成函数进行变换，从而得到新的数列。比方说，在生成函数上再乘以一个 x，我们就会让每一项的 x 的指数加 1，从而让整个



数列右移一位，得到了一个新的数列 F_{i-1}，即 0, 1, 1, 2, 3, 5, ⋯

$$\frac{x^2}{1-x-x^2}=x^2+x^3+2x^4+3x^5+5x^6+8x^7+13x^8+\cdots$$

现在，我们需要用各种代数运算手段，对等式左边的生成函数进行变换，让等式右边的展开式变成本文开头的那个数列。什么操作能够同时让数列第 1 项除以 2，第 2 项除以 4，第 3 项除以 8，以此类推，让所有的第 i 项都除以 2^i 呢？我们可以把所有的 x 都用 $\frac{x}{2}$ 来替代：

$$\frac{\left(\frac{x}{2}\right)^2}{1-\frac{x}{2}-\left(\frac{x}{2}\right)^2}=\frac{x^2}{4}+\frac{x^3}{8}+\frac{2x^4}{16}+\frac{3x^5}{32}+\frac{5x^6}{64}+\frac{8x^7}{128}+\frac{13x^8}{256}+\cdots$$

化简一下：

$$\frac{x^2}{4-2x-x^2}=\frac{x^2}{4}+\frac{x^3}{8}+\frac{2x^4}{16}+\frac{3x^5}{32}+\frac{5x^6}{64}+\frac{8x^7}{128}+\frac{13x^8}{256}+\cdots$$

这就是数列 $\frac{F_{i-1}}{2^i}$ 的生成函数了。接下来，我们想要让第 i 项系数乘以一个 i，也就是想要让每一项的系数都乘以该项的次数，这该怎么办呢？最神奇的地方出现了——我们对生成函数进行求导：

$$\left(\frac{x^2}{4-2x-x^2}\right)'=2\times\frac{x}{4}+3\times\frac{x^2}{8}+4\times\frac{2x^3}{16}+5\times\frac{3x^4}{32}+6\times\frac{5x^5}{64}+7\times\frac{8x^6}{128}+8\times\frac{13x^7}{256}+\cdots$$

也就是：

$$-\frac{(-2-2x)x^2}{(4-2x-x^2)^2}+\frac{2x}{4-2x-x^2}=2\times\frac{x}{4}+3\times\frac{x^2}{8}+4\times\frac{2x^3}{16}+5\times\frac{3x^4}{32}+6\times\frac{5x^5}{64}+7\times\frac{8x^6}{128}+8\times\frac{13x^7}{256}+\cdots$$

不过，求导的同时，x 的次数也移动了一位。我们在生成函数上再乘以 x，把 x 的次数纠正回来：

$$\left(-\frac{(-2-2x)x^2}{(4-2x-x^2)^2}+\frac{2x}{4-2x-x^2}\right)x=2\times\frac{x^2}{4}+3\times\frac{x^3}{8}+4\times\frac{2x^4}{16}+5\times\frac{3x^5}{32}+6\times\frac{5x^6}{64}+7\times\frac{8x^7}{128}+8\times\frac{13x^8}{256}+\cdots$$

这就是本文最初的那个数列的生成函数了。令 $x=1$，便有：

$$6=2\times\frac{1}{4}+3\times\frac{1}{8}+4\times\frac{2}{16}+5\times\frac{3}{32}+6\times\frac{5}{64}+7\times\frac{8}{128}+\cdots$$

答案跃然纸上！



36. 利用赌博求解数学问题

前一节的问题有一个非常自然的扩展：平均需要抛掷多少次硬币，才会首次出现连续的 n 个正面？显然，当 n 更大的时候，平均次数的计算将会更加复杂，很快就会演变成一大堆难以化简的代数式。有趣的是，这个问题的最终结论却出人意料地简单：为了得到连续 n 个正面，平均需要抛掷 $2^{n+1}-2$ 次硬币。简单美妙的结论让我们不由得开始思考，这个问题有没有什么可以避免复杂计算的巧妙思路？万万没有想到的是，在赌博问题的研究中，各种数学工具帮了不少大忙；而这一回，该轮到赌博问题反过来立功了。

设想有这么一家赌场，赌场里只有一个游戏：猜正反。游戏规则很简单，玩家下注 x 元钱，赌正面或者反面；然后庄家抛出硬币，如果玩家猜错了他就会输掉这 x 元，如果玩家猜对了他将得到 $2x$ 元的回报（也就是净赚 x 元）。

让我们假设每一回合开始之前，都会有一个新的玩家加入游戏，与仍然在场的玩家们一同赌博。每个玩家最初都只有 1 元钱，并且他们的策略也都是相同的：每回都把当前身上的所有钱都押在正面上。运气好的话，从加入游戏开始，庄家抛掷出来的硬币一直是正面，这个玩家就会一直赢钱。但只要输了一次，玩家就会把身上的所有钱全部输光，并失去继续参与游戏的资格。我们假设，赌场老板有一个心理承受能力极限：一旦有人连赢 n 次，赌场老板便会下令停止游戏，关闭赌场。让我们来看看，在这场游戏中存在哪些有趣的结论。

首先，连续 n 次正面朝上的概率虽然很小，但确实是有可能发生的，因此总有一个时候赌场将被关闭。赌场关闭之时，赚到钱的人就是赌场关闭前最后进来的那 n 个人。每个人都只花费了 1 元钱，但他们却赢得了不同数量的钱。其中，最后进来的人赢回了 2 元，倒数第二进来的人赢回了 4 元，倒数第 n 进来的人则赢得了 2^n 元（他就

是令赌场关闭的原因），他们一共赚取了 $2+4+8+\cdots+2^n=2^{n+1}-2$ 元。其余所有人初始时的 1 元钱都打水漂了，因为没有人挺过了倒数第 $n+1$ 轮游戏。

那么，整场游戏对玩家更有利还是对赌场老板更有利呢？或者换个问法，平均情况下，赌场老板是赚了还是亏了呢？答案当然是既不赚也不亏。由于这个游戏是一个完全公平的游戏，因此平均情况下，赌场的盈亏应该是平衡的，赌场老板的停业措施并不能改变这一点。因此，有多少钱流出了赌场，平均就有多少钱流进赌场。既然赌场被赢走了 $2^{n+1}-2$ 元，因此赌场的期望收入也就是 $2^{n+1}-2$ 元。而赌场收入的唯一来源是每人 1 元的初始赌金，这就表明游戏者的期望数量是 $2^{n+1}-2$ 个。换句话说，游戏平均进行了 $2^{n+1}-2$ 次。再换句话说，平均抛掷 $2^{n+1}-2$ 次硬币才会出现 n 连正的情况。



37. 非构造性证明

一个无理数的无理数次方有可能是一个有理数吗？

答案是肯定的。考虑$\left(\sqrt{2}\right)^{\sqrt{2}}$，如果它是一个有理数，问题就已经解决了。如果它不是一个有理数，那么$\left(\left(\sqrt{2}\right)^{\sqrt{2}}\right)^{\sqrt{2}}=\left(\sqrt{2}\right)^{\sqrt{2}\times\sqrt{2}}=\left(\sqrt{2}\right)^{2}=2$就是一个有理数。无论如何，我们都能找到一个无理数的无理数次方是有理数的例子。

这是非构造性证明的一个经典例子。我们虽然证明了存在两个无理数a和b，使得a^b是有理数，但我们却无法给出一组a和b的具体值来。毕竟我们也不知道，事实究竟是上述推理中的哪一种情况。

让我们来看一个更加典型的非构造性证明吧。我们把某个集合X的若干子集所组成的集合叫做X上的一个集族。考虑集合X上的一个集族，集族中的所有集合大小均为d。如果我们对X中的元素进行适当的红蓝二着色之后，每个集合里面都含有两种不同颜色的元素，我们就说这个集族是可以二染色的。例如，当$d=3$时，$\{1,2,3\}$、$\{1,2,4\}$、$\{1,3,4\}$、$\{2,3,5\}$就是可二染色的，把1、2染成红色，把3、4、5染成蓝色，则每个集合里都含有两种颜色。是否存在$d=3$时的不可二染色集族呢？当然存在。注意到，不管怎么对集合$\{1,2,3,4,5\}$中的元素进行二染色，我们总能找出三个颜色相同的元素，因此取集合$\{1,2,3,4,5\}$的全部（$C_5^3=10$个）元素个数为3的子集，总会有一个子集里面全是一种颜色。上述推理立即告诉我们，对于一个给定的d，一定存在一个集合个数为C_{2d-1}^{d}的不可二染色集族。这个数目还能再少吗？我们想知道，不可二染色集族中的集合个数最少可以少到什么地步。一个极其简单的证明给出了一个下界：集族的大小一定大于2^{d-1}。换句话说，对于任意一个集合个数不超过2^{d-1}的集族，一定存在一个二染色方案。



为了证明这一点，我们对 X 中的所有元素进行随机着色，每个元素被染成红色和蓝色的概率均等。那么，一个元素个数为 d 的集合中，所有元素均为同一种颜色的概率就应该是 $\frac{1}{2^{d-1}}$。如果集族内的集合个数只有不到 2^{d-1} 个，那么即使“各个集合中是否只含一种颜色”是互相独立的事件，这些事件的并集（即至少有一个集合内只含一种颜色）的概率也不超过 $\frac{1}{2^{d-1}}\times 2^{d-1}=1$，何况这些事件还不是独立的，因此存在单色集合的概率必然小于 1。这个概率值小于 1 说明什么？这说明，“至少有一个单色集合”并不是必然事件，一定有一种染色方案使得每个集合里都含两种颜色，换句话说就是该集族可以被二染色。

注意，我们用概率方法，证明了一个非概率型的事实！由此带来的另一个结果是，这是一个非构造性的证明。当 $d=3$ 时，这个定理告诉我们，任意一个只含 4 个集合的集族一定能被二染色。不过，我们只知道二染色的方案是存在的，却并不能给出一个具体的方案来。我们虽然证明了二染色方案一定存在，但证明过程却不能对我们寻找具体的方案给出任何提示。这就是非构造性证明的神奇之处。

在博弈论中，我们也有一些非构造性的证明。我们可以证明在某个游戏中，某位玩家有必胜的策略，但证明过程却不能告诉你，这个必胜策略究竟是什么。

记得我小学时就见过一个经典的奥数题目。两个人轮流在黑板上写一个不大于 10 的正整数。规定不准把已经写过的数的约数再写出来。谁最后没写的了谁就输了。显然，这个游戏是没有平局的，即使双方在每一步都使出最优策略，最终也还是有一个人会赢有一个人会输。也就是说，在这个游戏中，有一方玩家是可以必胜的。问是先写的人必胜还是后写的人必胜，必胜策略是什么？

答案很巧妙。先写者有必胜策略。他可以先写下数字 6，现在就只剩下 4、5、7、8、9、10 可以写了。把剩下的这 6 个数分成三对，分别是 $(4,5)$、$(7,9)$、$(8,10)$，每一对里的两个数都不成倍数关系，且 8 和 10 各自的约数恰好也在同一对里。因此不管你写什么数，我就写它所在的数对里的另一个数，这样可以保证我总有写的。

这个问题有一个很自然的扩展：规则不变，可以写的数扩展到所有不大于 n 的正整数。对于哪些 n 先写者必胜？证明你的结论。

其实，不管 n 是多少，先写者总有必胜策略。这时，就该轮到非构造性证明出场



了。考虑一个新的规则“不准写数字 1”。如果加上这个新规则后先写者有必胜策略，那么这个策略对于原游戏同样适用（因为 1 是所有数的约数，在先写者写完第一个数后，1 本来就不能写了）；如果在新规则下后写者必胜，则原游戏中的先写者一开始就把数字 1 写在黑板上，然后他就变成了新规则下的后写者。于是不管怎么样，先写者总是有必胜策略。

这种博弈游戏的分析技巧叫做“策略偷换”（strategy-stealing）。它的另一个经典例子是 Chomp 游戏。游戏在一块矩形的巧克力上进行，巧克力被分为 $M \times N$ 块。两人轮流选择其中一个格子，然后吃掉这一格及它右边、下边和右下角的所有格子。最左上角的那一块巧克力有毒，谁吃到谁就输了。图 1 是一个可能的对战过程。我们可以用类似的方法证明先手必胜。假设后手有必胜策略，那么先手把最右下角的那一块取走。注意到接下来对方不管走哪一步，最右下角的那一块本来也会被取走，因此整个棋局并无变化，只是现在的先手扮演了后手的角色，可以用后手的那个必胜策略来应对棋局，这样便巧妙地“偷”走了后手的必胜策略。

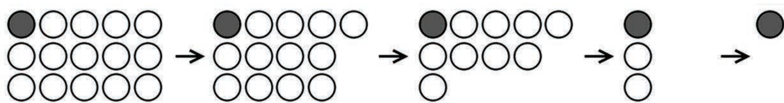


图 1

对于一些可能出现平局的游戏，我们也可以用类似的方法证明后手不可能有必胜策略。比如对于五子棋游戏（假设棋盘大小有限，并且没有禁手等规则），假设后手有必胜策略，那先手就随便走一步，以后就装成是后手来应对。如果在哪一步需要先手在已经下过子的地方落子，他就再随便走一步就是了。这样一来，先手便偷走了后手的必胜策略了，反而成了必胜的一方。这就说明，后手是不可能有必胜策略的。这种证明方法成立的前提就是，多走一步肯定不是坏事。事实上，对于所有这种“多走一步肯定不是坏事”的且信息透明决策对称的游戏，我们都可以证明后手是没有必胜策略的。既然后手没法必胜，那么我们立即可知：先手是一定有不败策略的。

不过，再次回到本节最开始的话题：我们虽然证明了谁有必胜策略或者谁有不败策略，但我们完全不知道具体的策略是什么！

38. 小合集（三）：数字问题

在前两个证明小合集中，我们看到了大量精彩的几何问题，以及几何图形在证明中的应用。在本部分的最后一节，让我们来欣赏一些数字问题吧。这些问题的证明大多是构造性的，其中有些构造非常大胆，同样令人叹服。

有两个数字类的问题我特别喜欢。先从第一个问题说起吧。

2 的 5 倍是 10，3 的 37 倍是 111，4 的 25 倍是 100……是否对于任意正整数 n，都能找到一个 n 的倍数，它全由数字 0 和 1 构成？

答案是肯定的。它的证明过程如下。

考虑数列 1, 11, 111, 1111，…。由于一个正整数除以 n 的余数只有 n 种可能，因此数列的前 $n+1$ 项中一定有两项，它们除以 n 的余数相同。这两项的差即满足条件。

在很多数字问题中，我们都会用到这样的思路。再看下面这个问题。

求证：存在一个正整数 n，使得 3^n 的末三位是 001。

证明如下。

由于 3 的幂有无限多个，但末三位只有 1000 种情况，因此我们一定能找到两个数 3^p 和 3^q，使得它们的末三位一样。不妨假设 $p>q$，于是 3^p-3^q 能被 1000 整除，即 $3^q(3^{p-q}-1)$ 能被 1000 整除。然而，3^p 与 1000 没有公因数，因此 $3^{p-q}-1$ 一定能被 1000 整除。也就是说，3^{p-q} 的末三位是 001。

这让我想起了下面这个问题。

不用计算，写出 99 999 的平方的末五位。



利用平方差公式可以很快得出答案：

由于 $99\,999^2-1$ 可以写成 $(99\,999+1)(99\,999-1)$，可见 $99\,999^2-1$ 能被 100 000 整除，也就是说 $99\,999^2-1$ 的末五位一定是 00 000，因此 $99\,999^2$ 的末五位一定是 00 001。

和大家分享一个非常捉弄人的题目：

899 是质数吗？

答案同样会用到平方差公式：

不是。因为 $899=900-1=30^2-1^2=(30+1)(30-1)=31\times29$。

这题妙就妙在，31 和 29 正好都是质数！如果用试除法，你必须要验证到最后才能得出正确结论。

初中数学竞赛中有一道与此相关的经典题目。

求证：当 n 是大于 1 的正整数时，n^4+4 一定不是质数。

答案如下：

$n^4+4=n^4+4n^2+4-4n^2=\left(n^2+2\right)^2-\left(2n\right)^2=\left(n^2+2+2n\right)\left(n^2+2-2n\right)$，由于 n 是大于 1 的正整数，易证最后所得的两个乘数也都是大于 1 的正整数。因此，n^4+4 总能分解成两个大于 1 的正整数之积，即它一定不是质数。

另一个类似的问题则是：

求证：若 n 是正整数，则 n^2+n+1 一定不是完全平方数。

证明非常简单，继续往下看之前你不妨先想一想。

因为 $n^2<n^2+n+1<n^2+2n+1=(n+1)^2$，说明 n^2+n+1 严格地介于两个相邻的完全平方数之间，因此它一定不可能是完全平方数。

哦，对了，说了这么久，我还没有提到第二个我最喜欢的数字问题呢。请看题目：

是否对于任意正整数 n，都能找到一个 n 的倍数，它含有从 0 到 9 所有的数字？

答案仍然是肯定的，它的证明更加简单。看了这个证明后，你一定会觉得自己笨死了。

假设 n 是一个 d 位数，那么 $1\,234\,567\,890\times10^d+1$ 和 $1\,234\,567\,890\times10^d+n$ 之间一定有一个数是 n 的倍数，它显然满足要求。

说到大胆疯狂的构造证明，不得不提到下面这个经典的定理：

证明，存在任意长的连续自然数序列，使得序列中的每一个数都是合数。

换句话说，相邻质数的间隔可以达到任意大。这个定理有如下这个极其简单的构造性证明。

任取一个正整数 $n>1$。由于 $n!$ 里含有因子 2，因此 $n!+2$ 仍然能被 2 整除。同理 $n!+3$ 能被 3 整除，$n!+4$ 能被 4 整除，等等，一直到 $n!+n$ 能被 n 整除。因此，$n!+2, n!+3, \cdots, n!+n$ 就是连续的 $n-1$ 个合数。由于 n 可以取到任意大，因此合数序列可以任意长。

另一个有趣的问题如下：

证明，对任意大的正整数 n，总能找到比 n 更大的三个正整数 a、b、c，满足 $a!b!=c!$。

也就是说，我们需要证明，$a!b!=c!$ 存在任意大的正整数解。答案如下：

注意到 $N\cdot(N-1)!=N!$。随便选一个正整数 $m>n$，令 $N=m!$，就有 $m!(m!-1)!=(m!)!$。

曾经见过下面这道更狠的数学竞赛题。

证明或推翻：$a^3+b^4=c^5$ 没有正整数解。

答案只有一句话，不知会让多少考生崩溃掉。

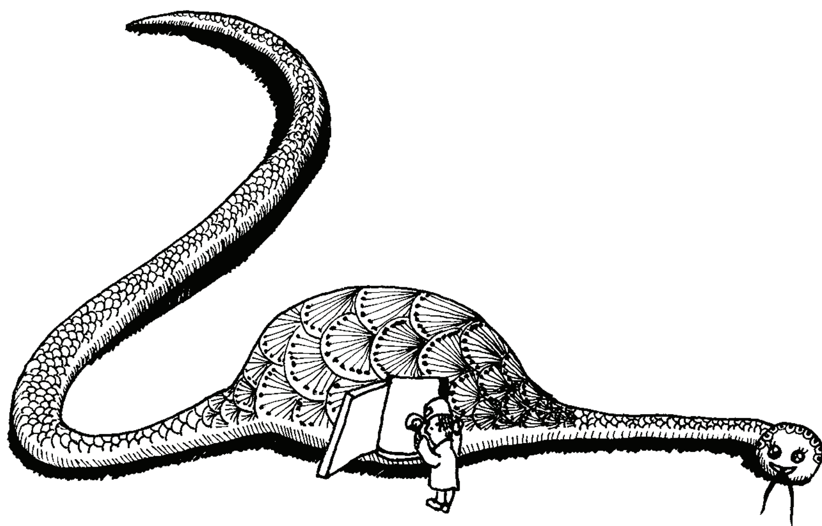
原命题是错误的。由于 $2^{24}+2^{24}=2^{25}$，因此 $\left(2^8\right)^3+\left(2^6\right)^4=\left(2^5\right)^5$。



第五部分

思维的尺度

如果你喜欢上一章最后一节那些宏伟的构造性证明，你一定会喜欢这一章。在这一章中，我们将会看到一些更加壮观的数学构造。即使是整个宇宙也无法超越人类思维的尺度。一道看似简单的数学问题，有可能会瞬间导出一个比宇宙中所有原子的数量更大的数。



39. 史诗般壮观的数学证明

你认为，是否有可能把全体正整数染成红蓝二色，使得不存在无穷长的等差数列，满足数列中的所有数都是一种颜色？

事实上，满足题意的染色方案是存在的。例如，我们可以从数字 1 开始，把正整数染成一段红一段蓝，使得每一段的长度都是其前一段的两倍。也就是说，我们把 1 染成红色，2 和 3 染成蓝色，4 到 7 染成红色，8 到 15 染成蓝色，依此类推，单色区间的长度成倍地增加。可以证明，如此染色后，一定不存在无穷长的单色等差数列。这是因为，假设某个等差数列的公差为 d，那么当单色区间的长度大于公差 d 时，等差数列将会一截一截地落在不同的染色区间中，从而拥有不同的颜色。

有趣的是，把上述命题中的“无穷长”换成“任意长”，命题就不见得仍然正确了。1927 年，荷兰数学家范・德・瓦尔登（Van der Waerden）证明了这么一个事实：对于任意大的正整数 k，总存在正整数 N，使得对从 1 到 N 的正整数进行红蓝二染色后，不管你是怎么染色的，里面总存在一个单色的长度为 k 的等差数列。当命题从两种颜色扩展到任意多种颜色时，该命题竟然也都成立。这个定理就叫做范德瓦尔登定理。下面，让我们来看看范德瓦尔登定理的证明过程。到了整个证明的最后一步，你一定会被震撼得说不出话来。

我们首先来证明 $k=3$ 的情况：存在某个 N 使得对 N 个连续自然数进行二染色后，里面总存在长为 3 的单色等差数列。我们把全体正整数 5 个 5 个地分组，1 到 5 为第一组，6 到 10 为第二组，以此类推。每一组里总共有 2^5 种不同的染色方案，因此在前 2^5+1 组里面必然有两个组的染色一模一样，比方说第 7 组和第 23 组吧。把第 7 组里的数分别记作 $A_1, A_2, \cdots, A_5$，由鸽笼原理，A_1、A_2、A_3 里面一定存在两个颜色相同的数，不妨假设 A_1 和 A_3 都是红色吧。考虑 A_5 的颜色：如果它是红色，我们的问题就



解决了，A_1,A_3,A_5 已经是一个长度为 3 的等差数列。下面考虑 A_5 是蓝色的情况。别忘了第 7 组和第 23 组的染色完全相同,因此如果把第 23 组里的数分别记作 $B_1,B_2,\cdots,B_5$，那么 B_1 和 B_3 也是红色，B_5 也是蓝色。下面我们来考虑全体正整数的第 39 组（注意到 7,23,39 是等差数列），我们把它里面的 5 个数分别记作 $C_1,C_2,\cdots,C_5$。考虑 C_5 的颜色：如果它是红色，那 A_1,B_3,C_5 就是一个全为红色的等差数列，否则 A_5,B_5,C_5 就是一个全为蓝色的等差数列。显然，上述证明中的“最坏情况”就是，第 1 组和第 33 组的染色相同，且第 1 组的第 1 个数和第 33 组的第 3 个数是红色的，则下一个数最远可以是第 65 组的第 5 个数，即全体正整数的第 325 个数。换句话说，$k=3$ 时 $N=325$ 即满足条件。（这不一定是最小的 N，但确实是一个满足要求的 N。）

有意思的是，对任意的颜色数 c，上述证明都是适用的。比方说，当 $c=3$ 时，取 $n=7\times\left(2\times3^7+1\right)$，把全体正整数 n 个 n 个分为大组，在头 3^n+1 组中必然存在两个染色一样的组，不妨把它们记作大组 A 和大组 B。把这两个大组里的数分别都 7 个 7 个地分成 $2\times3^7+1$ 个小组，在头 3^7+1 组中，必然有两个小组的染色方案一模一样，比如小组 a 和小组 b。

在每一个小组的前 4 个数中，必然有 2 个数的颜色是相同的，假设第 1 个数和第 4 个数是红色。那这个小组里面的第 7 个数要么是红色（问题已解决），要么是另一种颜色（比如蓝色）。将与前两个小组构成等差序列的第三个小组记作小组 c，由于一个大组中有 $2\times3^7+1$ 个小组，因此小组 c 一定还在该大组中。小组 c 中的第 7 个数要么是红色（问题已解决），要么是蓝色（问题已解决），要么是剩下的那种颜色（比如黄色）。那么，与两个大组构成等差序列的第三个大组（记作大组 C）中，对于相应的小组 c 位置上的第 7 个数（图 1 中标记星号的位置）的颜色就没有任何选择了，它或者和每个大组的那个染黄色的数成等差数列，或者和大组 A 中的小组 a 的蓝色数、大组 B 中的小组 b 的蓝色数构成等差数列，或者是跨越层数最多的，和大组 A 中的小组 a 的第 1 个红色数、大组 B 中的小组 b 的第 2 个红色数构成等差数列。

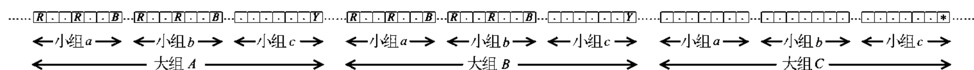


图 1

类似地，对于更大的颜色数 c，我们都可以归纳证明，只不过分组的层数更多，



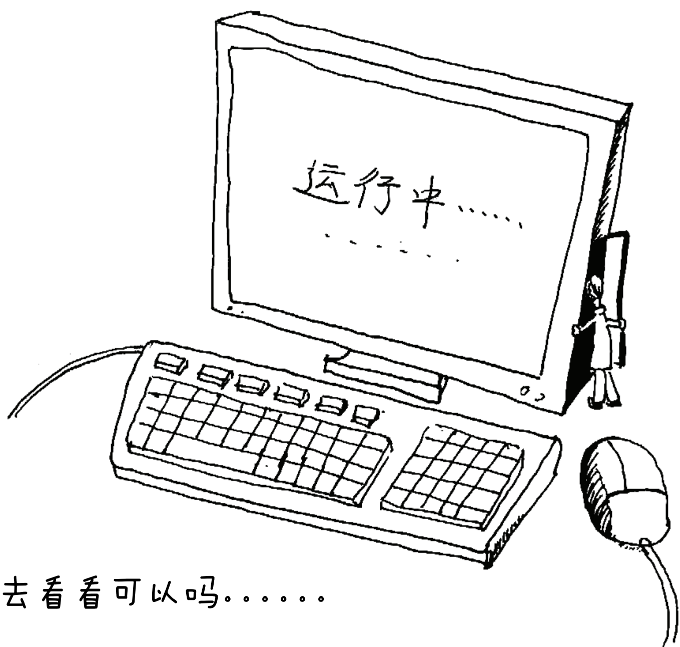
底数也相应变大。因此，我们解决了 $k = 3$ 且 c 任意大时的情形。

真正令人震撼的时刻到了。接下来，我们将对 k 施加归纳。下面尝试证明 $k = 4$ 、$c = 2$ 的情况，即存在一个充分大的 N ，使得对 1 到 N 的正整数进行二染色后，里面总存在长度为4的单色等差数列。由于当 $k = 3$ 时每325个数里面必然有一个等差数列，因此我们按每 487 个数一组进行分组。这样可以保证每一组里面的前 325 个数中总存在长为 3 的单色等差数列，并且该数列的第 4 个数也在该组内。注意，一个 487 元组共有 2^{487} 种染色方案，如果我们把它们看做 2^{487} 种不同的“广义颜色”，由 $k = 3$ 、$c = 2^{487}$ 的情况知，必然存在 3 个组，这 3 个组的位置形成等差数列，并且它们的染色方案完全相同。考虑每一组中前 325 个数所形成的长为 3 的等差数列，并考虑该数列中第 4 个数的颜色：如果颜色相同，问题解决；否则便考察顺推下去的第 4 个组的相应位置上的数的颜色，它将别无选择。

类似地，我们可以归纳出任意大的 k 和任意大的 c 的情形。可想而知，也可以说难以想象，用这种做法得出的 N 是何等地巨大，它将很快超出整个宇宙中任何具有实际意义的数字，其大小已经不能用通常的方式来记录了。这个证明的气势太宏大了，以至于当初很多人都没想到。当我第一次读到 2^{487} 种颜色时，视野一瞬间广大得难以描述；并且当我向着 k 更大的方向看去时，不禁对思维的尺度表示深深的膜拜。



40. 停机问题与“万能证明方法”



按照指定的语法规则编写代码，把你的想法“告诉”计算机；再在编译器中编译代码，生成一个个各有所能的程序。你可以随心所欲地制作各种搞怪的小程序：寻找并输出 2010 年到 2020 年中所有“黑色星期五”的日子，读取用户输入的手机号码并输出它是不是质数，读取用户输入的一篇文章并输出出现频率最高的四字词……这听上去似乎是一件相当有乐趣的事情。不过，程序员也有自己的苦恼。

最奇怪的幻想总是来自于最奇怪的需求。程序员一定有过这样的经历：看到自己编写的程序运行了半天都还没有任何结果，于是开始纠结，到底是再等一会儿呢，还



是强行终止程序，检查一下程序代码有没有写错。犹豫了半天决定杀掉进程，之后又检查了半天竟然发现程序没有写错。于是开始后悔，早知道程序没有死循环的话，刚才就多等一会儿了。此时，你会突然开始幻想，有没有什么编译器能够事先告诉你你的程序是否会无限运行下去？

注意，这里有一个假设：我们手里的计算机是一台理想的计算机。它拥有无穷无尽的内存，不会溢出，不会越界，可以存储任意大的、任意精确的数字。此时，“无限运行下去”就不仅仅是指“令 $a=1$ 并不断把 a 替换为 a^2 直到 $a>100$”这样的“死循环”了，状态永不重复的程序也有可能永远停不下来。比方说，“令 $a=100$ 并不断把 a 替换为 $a+1$ 直到 $a<10$”，这段程序虽然不会重复之前的状态，但也不会停下来。

在这样的假设下，编程判断一段代码是否会无限执行下去将会相当地困难。但我们仍然不排除会有某个天才程序员想出了超级复杂的算法，耗时五年为他心爱的编译器写出了这样一个强大的插件。为什么不可能呢？这个东西看上去似乎比时光旅行机更现实一些。或许我们会在某个科幻电影中看到，一个程序员在漆黑的屏幕上输入几个数，敲了一下回车，然后屏幕上立即用高亮加粗字体显示：“警告：该输入数据会导致程序无限运行下去，确定执行？(Y/N)”如果有一天，这一切真的成为了现实，那么你能利用这个玩意儿来做些什么实用的、有价值的事情？如果我说你能靠这玩意儿发大财的话，你相信吗？

永远不要说什么“看看当年我参加计算机竞赛时第二题的第四个数据点是不是真的因为死循环才超时的”、“看看上个星期做完项目跟客户演示时为什么半天没有输出”之类的话。如果是我的话，我一定会用点儿别人想不到的雕虫小技干出一番惊天动地的大事。我上来就先写一个哥德巴赫猜想的验证程序。我写一个程序，让它从小到大枚举所有的偶数，看是不是有两个质数加起来等于它。如果找到了，继续枚举下一个偶数，否则输出这个反例并结束程序。然后编译该程序。这个编译器不是可以预先判断我这个程序能否终止吗？如果编译器说我这个程序会无限执行下去的话，我岂不是相当于证实了哥德巴赫猜想吗？或者，编译器说程序最终会终止，那哥德巴赫猜想不就直接被推翻了吗？不管怎样，我都将成为解决哥德巴赫猜想的第一人，在数学史上留下自己的名字。接下来呢？把刚才的程序代码改成孪生素数搜索器，再用编译器检验一下，看看是不是真的有无穷多个孪生素数。梅森素数是否有无穷多个，这也是数



论中长期以来悬而未决的难题。不过现在看来，我也能不费吹灰之力就把它解决了。还记得我们在讲“最折磨人的数学未解之谜”时提到的 $3x+1$ 问题吗？写一个“证明程序”也只是几分钟的事情，而且还能拿走埃尔德什提供的 500 美元奖金呢。数学上的未解之谜多着呢，我永远不愁没事做。1984 年，马丁 · 拉巴尔（Martin LaBar）询问是否能用 9 个不同的平方数构成一个 3×3 的幻方，这个问题的奖金目前已经累积到了 100 美元加 100 欧元再加一瓶香槟。网上搜索“数学未解难题”，看看哪些问题是离散的，其中又有哪些问题是有悬赏的，写几个程序就可以把它们统统解决……

还是从幻想中清醒过来吧，判断一个程序能否无限运行下去的程序是否真的存在我们还不知道呢。不过仔细回想一下上面的讨论，你会意识到这种程序的存在该是多么不可思议。即使这样的程序真的存在，实现它的难度也绝对不亚于解决上述任何一个数学猜想，不然的话大家都转而向这个神奇的“万能证明方法”进攻了。

而事实上，计算机科学家们已经从理论上证明了，这种程序是永远不可能实现的。在计算机理论中，该结论可以叙述为：停机问题是一个不可解的问题。停机问题不可解的证明并不复杂，并且非常有趣。用反证法，假设我们有一个满足要求的程序 $P(a,b)$，它可以预先判断出运行代码 a 并读入数据 b 之后程序是否会终止。那么，我们可以编写这样一个程序 Q，它首先读取输入数据并把它记做 x，然后调用 $P(x,x)$ 并根据其返回的结果执行不同的任务：如果 $P(x,x)$ 返回的结果是“不会终止”，立即退出程序；否则，任意执行一个死循环任务，比如“令 $a=1$ 并不断把 a 替换为 a^2 直到 $a>100$”。现在，运行程序 Q，然后把程序 Q 本身的代码作为输入数据传进去，于是程序 Q 调用 $P(x,x)$ 时，实际上问的是“运行程序 Q 并输入 Q 的代码后会发生什么”，也就是询问它自身的命运。但根据程序 Q 的规则，如果 $P(x,x)$ 认为该程序不会终止，Q 就会立即退出；如果 $P(x,x)$ 认为该程序总有终止的一刻，程序 Q 反而陷入循环。于是，$P(x,x)$ 并没有成功预测此时 Q 的命运，这说明停机问题是永远无法解决的。

我们刚才那些美好的梦想被这一个简短的证明撕成了碎片。

永远不要小瞧人类的想象力。对“万能证明方法”的进攻并没有就此结束。虽然判断一段代码运行后是否会终止的程序是不存在的，但这个“万能证明方法”的思路是非常值得借鉴的。下面，我给你一个绝对存在的东西，它同样可以用于我们的“程序证明”。它就是指定的编程语言中任意一段代码运行后最终会停止下来的概率。假如

说这种编程语言有 p 种字符（包括代码结束的标识符共 $p+1$ 种），长度为 n 的代码中有 $T(n)$ 个不会无限运行下去，那么我们定义这个概率就是 $\sum_{n=1}^{\infty}\frac{T(n)}{(p+1)^n}$ 。考虑两种极端的情况：如果所有代码都永不终止，那么这个概率值为 0；如果所有代码运行后最终都能停下来（包括语法错误根本不能编译的情况），那么概率为 $\sum_{n=1}^{\infty}\frac{p^{n-1}}{(p+1)^n}$ ，其中 p^{n-1} 是除去那个结束符号后所有可能的 $n-1$ 位代码的数量。利用等比数列求和公式容易得出这个值等于 1。当然，在实际情况下，这个概率值是一个介于 0 和 1 之间的确定的数。

有了这个概率之后，我们就无敌了！我们可以故技重施，写一个哥德巴赫猜想验证程序。假如这个程序的长度为 L 。注意到 $\sum_{n=k+1}^{\infty}\frac{p^{n-1}}{(p+1)^n}=\left(\frac{p}{p+1}\right)^k$ ，当 k 足够大时必然有某个时刻 $\left(\frac{p}{p+1}\right)^k$ 比 $\frac{1}{(p+1)^L}$ 小。我们再用一个程序来生成所有可能的长度不超过 k 的代码。然后便是壮观的一幕：让所有这 $\sum_{n=1}^{k}p^{n-1}$ 个程序同时运行！这里面，有些程序的代码语法有错，根本就不能通过编译；有些程序运行后屏幕一闪就退出来了；有些程序可能得等个好几天才能退出；当然也有将要无限运行下去的程序。不过可以肯定的是，受到上面那个概率值的限制，最终停止下来的程序是有一个上限的。随着越来越多的程序停止下来，总有某个时刻会达到这样一种状态：终止的程序所占的比例与我们那个概率值的误差不到 $\frac{1}{(p+1)^L}$ （因为我们没有考虑的那些代码即使全部都会终止也只占 $\left(\frac{p}{p+1}\right)^k$ 的分量，而 k 值的选择保证了这个分量不足 $\frac{1}{(p+1)^L}$ ），此时只要再有一个长度不超过 L 的程序终止，实际比例（将增加至少 $\frac{1}{(p+1)^L}$ ）就超过那个概率了。这时，我们就可以肯定，到时候还没有停止的程序必然将无限运行下去。如果届时我们的哥德巴赫猜想还没找出反例，这就意味着这个程序永远找不出反例了，哥德巴赫猜想也就得到了证实。

或许，这个工程确实有点庞大，需要耗费大量的时间和金钱。不过，为了证明那么多悬而未解的数学之谜，投入再多的时间和资金也值得啊！我们还可以采取分布式计算的办法，邀请全球的计算机一起来参与计算！那么，为什么不这样做呢？真实的



情况到底如何呢?

这或许会有些不合常理：我们上面提到的那个概率值是一个“不可计算数”（uncomputable number）。它是一个可以严格定义出来并且也确实存在的数，但我们永远无法计算出它的值（即不存在某种算法能够给出小数点后任意多位的数字）。这个概率值是有名字的，它叫做蔡廷常数，是以数学家和计算机科学家格里高里·蔡廷（Gregory Chaitin）的名字命名的。可以证明，蔡廷常数确实是不可计算的。不妨反过来想，假如我们有一个能够给出蔡廷常数小数点后任意多位的值的算法，那么我们就能用上面那种“等足够多的程序终止”的方法判断出一个代码长为 n 的程序是否会无限运行下去，这相当于有了一个解决停机问题的算法。但我们前面已经证明了，停机问题是不可解的，因此可以肯定地说，算出蔡廷常数一定是不可能的。

41. 奇怪的函数（一）

教高中数学竞赛辅导课时，我在某次课堂测验中出了这么一道题：

构造一个从全体正整数映射到全体正整数的函数 $f(n)$，使得每一个正整数都被映射过无穷多次。

乍一看，这似乎很难办到。我们可以令 $f(n)$ 等于 n 除以 1 000 000 的结果的整数部分，让每个正整数都被映射过 1 000 000 次；也可以令 $f(n)$ 等于 n 除以 1 000 000 的余数，让 1 000 000 以内的正整数都被映射过无穷多次。不过，我们真的能让所有正整数都被映射无穷多次吗？

答案是肯定的。当时，我自己提供的标准答案是：

令 $f(n)$ 等于 n 的各位数字之和。

例如，808 067 的各位数字之和是 $8+0+8+0+6+7=29$，因此 $f(808\,067)=29$。很容易看出，任意一个正整数都有无穷多个原象。比方说，对于某个正整数 m，令 n 为 m 个数字 1 相连组成的 m 位数，于是就有 $f(n)=m$。在 n 里面的任意位置添加任意多个 0，其函数值仍然为 m，因而 m 有无穷多个原象。

看到学生们交上来的试卷后，我非常高兴。绝大多数学生的答案都是正确的，并且他们的构造思路完全不同。某个学生写的是：

令 $f(n)$ 等于去掉 n 中的所有数字 9，把剩下的数当做九进制并将其转换为十进制后的结果。

例如，去掉 9 012 998 中的所有数字 9 后，得到一个新的数 128。把 128 当做九进制数，转换为十进制数后便是 107。于是 $f(9\,012\,998)=107$。不难看出，这个函数也



满足要求。

美中不足的是，这个函数的函数值有可能等于 0，而题目则要求函数的值域不包含 0。其实，这个问题不大，我们只需要在函数 $f(n)$ 的定义后面加一句“如果算出来的结果为 0，就随便取一个正整数（比如 1）作为函数值”或者“把算出来的结果再加 1 作为函数值”即可。

另一个学生则写道：

令 $f(n)$ 等于 n 中所含质因数 2 的个数。

例如，358 400 可以分解为 $2^{11}\times5^2\times7$ 。于是 $f(358\,400)=11$ 。显然，这也是一个满足要求的答案。注意，这个函数也存在上面提到的值域问题，不过也可以用类似的方法进行完善。

上述答案都很巧妙。不过，下面这个才是当时我所见到的最简单、最直接的答案：

令 $f(n)$ 等于数列 $1, 1, 2, 1, 2, 3, 1, 2, 3, 4, 1, 2, 3, 4, 5, \cdots$ 中的第 n 项。

当然，方法还有很多。你还能想出多少来？

42. 比无穷更大的无穷

对上一节中的函数稍作改造，我们还能得到更加违反直觉的函数。例如，我们可以构造一个从正整数到正有理数的一对一函数，从而说明正整数和正有理数一样多！

方法很简单。取 $f(n)$ 为上一节中任意一个把全体正整数映射到全体正整数，并且每一个正整数都被映射过无穷多次的函数。按照如下方式定义 $g(n)$：对于某个 n，如果 $f(n)$ 的函数值 m 是第 i 次被映射到，则令 $g(n)$ 等于分母为 m 的所有最简分数从小到大排列后的第 i 个分数。

比方说，我们取 $f(n)$ 为数列 1, 1, 2, 1, 2, 3, 1, 2, 3, 4, 1, 2, 3, 4, 5, ⋯ 中的第 n 项。由于 $f(8)$ 已经是第三次映射到 2 了，因此 $g(8)$ 就等于分母为 2 的第三个最简分数，即 $\frac{5}{2}$。

n	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$f(n)$	1	1	2	1	2	3	1	2	3	4	1	2	3	4	5
$g(n)$	$\frac{1}{1}$	$\frac{2}{1}$	$\frac{1}{2}$	$\frac{3}{1}$	$\frac{3}{2}$	$\frac{1}{3}$	$\frac{4}{1}$	$\frac{5}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{5}{1}$	$\frac{7}{2}$	$\frac{4}{3}$	$\frac{3}{4}$	$\frac{1}{5}$

于是，$g(n)$ 就是一个把全体正整数映射到全体正有理数的函数，并且每个正有理数都被映射且只被映射过一次。这意味着，全体正整数和全体正有理数之间存在一个一一对应的关系，有多少个正整数，就有多少个正有理数！

大家或许会觉得奇怪：正有理数集不但包含了正整数集的所有数，还包含了正整数集没有的数，这两个集合里的元素怎么可能一样多呢？不过，对于一个无穷集合来说，既无重复又无遗漏地映射到一个比自己大的集合，这不是什么稀罕的事。比如，即使乍看上去，正整数集比非负整数集少一个数字 0，但它们之间仍然存在一对一的函数。最简单的例子就是 $f(n)=n-1$。



n	1	2	3	4	5	6	7	8	9	10
$f(n)$	0	1	2	3	4	5	6	7	8	9

不仅如此，正整数集和全体整数集之间也能一一对应起来。比方说，规定当 n 为偶数时 $f(n)=\frac{n}{2}$，当 n 为奇数时 $f(n)=-\frac{n-1}{2}$，则有以下对应情况。

n	1	2	3	4	5	6	7	8	9	10
$f(n)$	0	1	−1	2	−2	3	−3	4	−4	5

注意到，我们之前已经有了一个从正整数到正有理数的一对一函数，因而自然也就有了负整数与负有理数之间一一对应的方法。现在，我们又有了一个从正整数到全体整数的一对一函数，其中全体整数里包括了正整数、0、负整数三个部分，它们各自可以和正有理数、0、负有理数形成一一对应。也就是说，我们可以立即构造出一个复合函数，它能把正整数集一一映射到全体有理数集上。因而，正整数和全体有理数也是一样多的！

可见，在比较无穷集合的大小时，违反直觉的现象太多了。我们必须要给无穷集合谁大谁小下一个严格的定义。

格奥尔格·康托尔（Georg Cantor）是伟大的德国数学家，集合论的创立者，勇敢正视无穷集合的第一人。他提出，在判断无穷集合大小时，我们不应该再着眼于狭义的“元素个数”，而应该采用一种类似于“集合大小规模”的概念。他把这个新的概念叫做“集合的势”。康托尔规定，只要我们有办法把两个集合中的元素一一对应起来，那么这两个集合的大小就是相等的，或者说它们是等势的。按照这种定义，不管是非负整数集，还是全体整数集，甚至是全体有理数集，都和正整数集等势，它们是一组规模相同的无穷集合。

一个集合与正整数集等势，意思就是这个集合中的元素与正整数之间存在一一对应的关系。换句话说，尽管这个集合中的元素有无穷多，但我们能按照某种方式对它们进行排序并编号，用“第一个元素是谁，第二个元素是谁，第三个元素又是谁”的方式把它们一一列举出来。因而，我们给所有与正整数集等势的集合取了一个形象的名字，叫做“可数集”。刚才讨论的非负整数集、全体整数集、全体有理数集都属于可数集的范畴。



像$\sqrt{2}$和$\frac{1+\sqrt{5}}{2}$这样的数构成的集合是否可数呢？虽然它们并不是有理数，但它们都是某个整系数多项式方程$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$的解。比方说，$\sqrt{2}$就是$x^2 - 2 = 0$的一个解，$\frac{1+\sqrt{5}}{2}$则是$x^2 - x - 1 = 0$的一个解。我们把所有这样的数叫做“代数数”。代数数不但包括了所有的有理数（因为有理数可以看做一次方程的解），还包括了像$\left(\frac{\sqrt{3}}{2}+\frac{\sqrt[3]{5}}{7}\right)\times 13\sqrt[100]{2}$一样的怪数，甚至包括了一些高次方程中无法用常规手段表达出来的解。有趣的是，即使是代数数这样庞大的数字群体，仍然属于可数集的范畴，因为我们可以按照一定的顺序把它们依次排列起来。

具体怎么做呢？首先，由于每个整系数多项式方程的解的个数都有限（不会超过这个多项式的次数），因此我们只需要找到一种排列这无穷多个多项式的方法即可。最容易想到的方案自然是，按照次数由低到高对多项式排序。这种方案可行吗？不行。$x+1=0$，$x+2=0$，$x+3=0$……都是一次多项式方程，这样的多项式方程有无穷多个。如果按照次数高低给多项式排序，二次多项式将永远也排不上号。

我们可以按照所有系数之和给多项式排序吗？也不行。因为多项式的系数有可能是负的，因而其系数之和可以任意小，根本不存在系数之和最小的多项式。

我们可以按照所有系数的绝对值之和给多项式排序吗？这次似乎有些希望，但细想一下你会发现还是不行。$x-2=0$，$x^2-2=0$，$x^3-2=0$……它们的系数绝对值之和都是 3。这样的多项式方程有无穷多，并且它们对应着不同的代数数。那些系数绝对值之和更大的多项式，就永远也排不上号了。

我们来总结一下之前种种方案失败的原因：如果只限定次数，系数将有无穷多种选择；如果只限定系数，次数将有无穷多种选择。如果同时对次数和系数加以限定，问题不就解决了吗？因此，我们想到，为何不在所有系数的绝对值之和的基础上，再加上这个多项式的次数，按照这个结果的大小给多项式排序呢？具体地说，定义n次整系数多项式$a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$的“复杂程度”为$n+|a_0|+|a_1|+\cdots+|a_{n-1}|+|a_n|$，那么对于任意一个正整数$N$，复杂程度恰好为$N$的整系数多项式都只有有限多个。这保证了我们可以顺利地给多项式排序。

下面就是一种给所有整系数多项式方程排序的方法。



(1) 总方针：按照多项式的复杂程度值从小到大排序。

(2) 如果复杂程度相同，则按照次数由低到高排序。

(3) 如果次数也相同，则按照常数项由小到大排序。

(4) 如果常数项也相同，则按照一次项系数由小到大排序。

(5) 如果一次项系数也相同，则按照二次项系数由小到大排序。

……

现在，我们已经给所有整系数多项式方程找到了一个合适的排列顺序，让每个整系数多项式都有了一个编号。从小到大依次写出第一个方程的解，再依次写出第二个方程的解，再依次写出第三个方程的解……这样便能穷举完所有的代数数。最后，把重复出现的数删去，从而得到了目标数列——每个代数数都恰好出现了一次。代数数与正整数之间也就有了一一对应的关系。或者说，代数数也是可数的。

事实上，建立代数数与正整数的一一对应关系远没有那么复杂。我们还有一种更帅的方法。如果把一个整系数多项式方程中所有的符号都列出来，一共也就只有0、1、2、3、4、5、6、7、8、9、+、−、=、x 共 14 种符号。多项式方程 $x^2-2=0$ 也就可以简单地看做由符号 x、2、−、2、=、0 相连得到的。这样看来，给多项式排序就变得异常简单了。

(1) 总方针：按照多项式所含符号个数从少到多排序。

(2) 如果多项式所含符号个数相同，按照第一个符号排序。

(3) 如果第一个符号也相同，按照第二个符号排序。

(4) 如果第二个符号也相同，按照第三个符号排序。

……

同样地，依次写出每个多项式方程的每一个解，再去掉重复的数，我们将得到把正整数一一映射到全体代数数的另一种方案。

那么，数学世界中最为神秘的数学常数 π 呢？直觉上，某个超级复杂的整系数多项式方程，解出来恰好是 $x=\pi$，这是绝对不可能的。正如我们在第 19 节所述，1882 年，德国数学家林德曼证明了，π 确实不满足任何整系数多项式方程，即 π 不是代数数。$e\approx 2.718$ 则是另一个重要的数学常数，它也不是代数数。这是由数学家查尔



斯・埃尔米特（Charles Hermite）在 1873 年首次证明的。我们把这些不是代数数的数都叫做“超越数”。

虽然 π 和 e 属于更没规律的超越数，但我们仍能把它算出来。我们可以设计出一套算法，只要给它足够长的时间，它就能计算出 π 或者 e 的小数点后任意多位数。我们把这种能够计算到任意精度的数叫做“可计算数”。按照这种定义，整数、有理数、代数数都是可计算数，除此之外，π、e 也是可计算数。人们目前尚不知道 $\pi+e$、$\pi-e$、πe、$\frac{\pi}{e}$、π^e 是不是超越数（事实上，人们现在还不知道它们是不是无理数），但不管怎样，它们也都是可计算数。可见，可计算数集合的范围远远超过了之前讲到的整数集、有理数集以及代数数集。然而，下面我们将证明，可计算数仍然是可数的。

借用代数数可数性的第二种证明方法，我们可以很快给全体可计算数制定一个排序方案。我们可以选用某种计算机编程语言来描述可计算数的计算方法，那么所谓的计算方法，说白了也就是由标准美式键盘上的 95 种字符（包括小写字母、大写字母、数字以及各种符号）组成的一段程序代码。我们把所有可能的代码按照代码长度排序，长度相同者按第一个字符排序，第一个字符也相同则按第二个字符排序，以此类推。这将列举出所有可能的程序代码，从而也将列举出所有可能的可计算数。因此，可计算数集合也是可数的。

有没有什么数是不可计算的呢？答案是肯定的。在第 40 节的末尾，我们讲到了蔡廷常数，它就是一个典型的不可计算数。不过，它虽然不可计算，但却有一个明确的定义，可以用数学语言清晰地定义出来。我们把所有这样的数都叫做“可定义数”。

可定义数不但包含我们平常经常使用的整数和有理数，也包含代数数、可计算数，甚至还包含一些不可计算数（例如蔡廷常数）。事实上，历史上一切被数学家研究过的数，不管是已发表的还是未发表的，不管是已命名的还是未命名的，不管是能算出来的还是不能算出来的，都是可定义数。这是一个异常庞大的数集。不过，它仍然是可数的。

为了证明这一点，只需要注意到，一个数的定义，无非是用数字、字母和标点符号组成的一段英文文本。我们把所有能够用来定义一个数的英文文本找出来，按照文本的长度由短到长排序（长度相同者按照之前的方法处理）。因此，所有可定义数组



成的集合也是可数的，可定义数和正整数之间存在一一对应的关系。

看到这里，想必大家会有一个疑问：讲了半天，究竟有没有不可数的无穷集合呢？换句话说，我们能找到比正整数集规模更大的无穷吗？

答案是肯定的。康托尔发现，不可数的集合是存在的。例如，所有介于 0 和 1 之间的实数就是不可数的。你不能从小到大对 (0,1) 区间内的所有实数进行排序，因为不存在“第一个正实数”。你也不能按照小数展开的长度对所有实数进行排序，因为很多数都是无限小数，它们永远排不上号。事实上，康托尔用一种非常巧妙的方法证明了，任何方案都不能把正整数和 (0,1) 区间内的实数一一对应起来。

证明的思路是，先假设我们已经按照某种方案将 (0,1) 区间内的所有实数列成了一张表，然后再说明，这张表其实并没有包含所有的实数。

现在，假设我们已经把所有 (0,1) 区间内的实数按照某种顺序排列为 $a_1, a_2, a_3, \cdots$。这里面的每个数都可以表达为形如“零点几几几几几……”的无限小数（如果是有限小数，可以在其后面添加数字 0，把它变成无限小数）：

$a_1 = 0.314\ 159\ 265\ 3\cdots$

$a_2 = 0.808\ 080\ 808\ 0\cdots$

$a_3 = 0.670\ 000\ 000\ 0\cdots$

$a_4 = 0.222\ 222\ 222\ 2\cdots$

$a_5 = 0.618\ 033\ 988\ 7\cdots$

$a_6 = 0.123\ 434\ 343\ 4\cdots$

……

下面我们构造一个新的实数，它也属于 (0,1) 区间，但却不在这张列表里。让这个实数的小数点后第一位不等于 a_1 的第一位，第二位不等于 a_2 的第二位，等等，总之要让这个实数的小数点后第 n 位不等于 a_n 的第 n 位。那么，这个新实数将有别于上面那个列表中的任何一个数，因为它和列表里的任意一个数都有至少一位是不同的。因此，我们永远不可能把所有 0 到 1 之间的实数一个也不少地排成一列。

注意，在上述证明过程中，“实数区间”这个条件用在了哪里。我们当然也可以假设 (0,1) 区间内的有理数被排成了序列 $a_1, a_2, a_3, \cdots$，也可以像刚才那样构造出一个数，它不等于序列中任何一个数。不过，我们构造出的这个数不见得仍然是有理数，



因此这和假设并不矛盾。但在证明实数区间不可数时，我们构造出来的数的确是本应该在列表中的数，这才是“实数区间”这个条件发挥作用的地方。

这是证明 (0,1) 区间内所有实数不可数最传统、最经典的方法。不过，作为实数理论中的一个基本结论，它还有很多不同证明。我想在这里介绍另一种同样漂亮的证明，这是由马修・H. 贝克（Matthew H. Baker）在 2008 年提出的。

设想 A 和 B 两个人在实数区间 (0,1) 上玩一个游戏。首先，A 在 (0,1) 区间选一个数 a_1，然后 B 在区间 $(a_1,1)$ 里选一个数 b_1。接着，A 在 (a_1,b_1) 区间选一个数 a_2，然后 B 在区间 (a_2,b_1) 里选一个数 b_2……总之，A 只能选取越来越大但不能比 B 选的数更大的数，B 只能选取越来越小但不能比 A 选的数更小的数。两人轮流按规则选数，这些数将从两侧出发不断向中间靠拢。可以看到，序列 $a_1,a_2,a_3,\cdots$ 是一个单调递增的有界序列，因此游戏无限进行下去，序列 $a_1,a_2,a_3,\cdots$ 最终会收敛到某一个实数 c。游戏进行前，A 和 B 约定区间 (0,1) 的一个子集 S，规定如果最后 c 在 S 中，A 胜，否则 B 胜。

一个有趣的事实是，如果 S 是可数集，B 肯定有必胜策略。如果集合 S 是可数的，那么 B 就可以把集合 S 里的数排列成一个序列 $s_1,s_2,s_3,\cdots$。B 的目标就是让序列 $a_1,a_2,a_3,\cdots$ 的极限不等于序列 $s_1,s_2,s_3,\cdots$ 中的任一个数。考虑 B 的这样一个游戏策略：当 B 第 i 次选数时，如果选 s_i 合法，那么就选它（这样序列 $a_1,a_2,a_3,\cdots$ 就不能收敛到它了）；如果选 s_i 不合法，那就随便选一个合法的数（反正序列 $a_1,a_2,a_3,\cdots$ 已经不可能收敛到 s_i 了）。这种策略就可以保证 A 选出的数列的极限不是集合 S 里的任一个数。

接下来就是神奇的一刻了。假如 A 和 B 约定好的集合 S 就是整个实数区间 (0,1)，那么 B 显然不可能获胜；但如果 (0,1) 是可数集，B 是有必胜策略的。于是我们就知道了，(0,1) 是不可数集。

既然连 (0,1) 区间都不可数，整个实数集 $\mathbb{R}$ 当然也就是不可数的了。有趣的是，我们可以在 (0,1) 区间和整个实数集 $\mathbb{R}$ 之间建立一一对应的关系，从而说明 (0,1) 区间和实数集 $\mathbb{R}$ 是同一级别的无穷集合，正如正整数集和有理数集是同一级别的无穷集合一样。比如函数 $f(x)=\tan\left(\pi\left(x-\frac{1}{2}\right)\right)$，它是一个从 (0,1) 区间到实数集 $\mathbb{R}$ 的一对一



函数，这就证明两个集合是等势的。利用图 1 所示的几何方法，我们也能马上看出，如果两根线条的长度不同，它们上面的点也能一一对应起来。即使其中一根线条无限长，我们同样能找到一一对应的方法。我们把所有和 (0,1) 区间等势的集合都叫做“$\mathcal{C}$ 势集”。

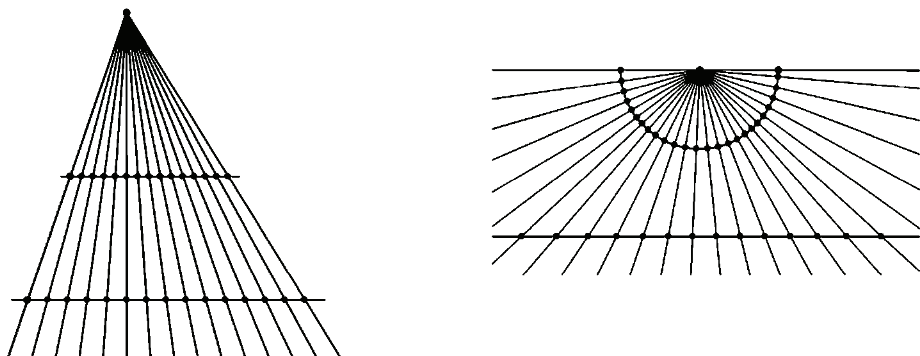


图 1

现在，我们已经有了两种规模不同的无穷：以正整数集为代表的可数集（包括整数集、有理数集、代数数集、可计算数集、可定义数集），以 (0,1) 区间为代表的 $\mathcal{C}$ 势集（包括任意长的连续实数区间，甚至是整个实数集 $\mathbb{R}$ ）。这是两个尺度完全不一样的无穷。与可数集比起来，$\mathcal{C}$ 势集真的是非常非常大的集合。

有理数集是可数的，但实数集是不可数的，这样我们便可得到一种无理数存在性的非构造性证明。事实上，在全体实数中，几乎所有的数都是像 $\sqrt{2}$ 一样的无理数，有理数仅仅是实数中微不足道的一部分。

类似地，由于代数数是可数的，而实数集不可数，因而我们可以立即推出超越数的存在性。事实上，几乎所有的数都是像 π 、e 那样的超越数，代数数只占实数集中零星的一部分。

同理，由于可计算数是可数的，而实数集不可数，因而实数集中必然存在不可计算数。事实上，几乎所有的数都是像蔡廷常数一样的不可计算数，相比之下，可计算数实在是少得可怜。

同理，由于可定义数是可数的，而实数集不可数，因而在实数集中一定存在不可

定义数。

那么，在数学历史上，谁发现了第一个不可定义数呢？答案是，从没有人发现过不可定义的数，以后也不会有人找到不可定义的数。因为不可定义数是无法用语言描述的，我们只能用非构造的方式证明不可定义数存在，但却永远没法找出一个具体例子来。

在实数当中，几乎所有的数都是不可定义的。数学家们所研究的数，只是实数世界中的沧海一粟。不过，数学家们也不会损失什么。每一个值得研究的数一定都有着优雅漂亮的性质，这些性质就已经让它成为了能够被定义出来的数。



43. 奇怪的函数（二）

在高中时代，我就已经有收集“另类函数”的爱好了。学习周期函数时，老师告诉我们，常函数（比如 $f(x)=1$ ）也是周期函数，只不过它们比较特殊——没有最小正周期。当时我就在想，除了常函数以外，还有没有其他的没有最小正周期的周期函数呢？某次看书时，我意外地发现，竟然真的有这样的函数。考虑这么一个函数 $f(x)$：当 x 是有理数时，函数值为 1；当 x 是无理数时，函数值为 0。由于对于任意一个有理数 q，都满足有理数加上 q 还是有理数，无理数加上 q 还是无理数，因此一切有理数 q 都是这个函数的一个周期。由于不存在最小的正有理数，因而这个函数也就没有最小正周期。

后来我才知道，这个函数叫做狄利克雷（Dirichlet）函数，它是数学分析中非常经典的异形函数。它拥有大量违背直觉的性质，给很多看似成立的数学命题提供了反例。例如，狄利克雷函数竟是一个处处不连续的函数！对狄利克雷函数稍作修改，我们还可以构造出乍看上去更加不可思议的函数。例如，定义这样一个函数 $f(x)$：当 x 是有理数时，函数值就取 x 本身；当 x 是无理数时，函数值为 0。利用函数连续性的定义不难证明，这个函数只在 $x=0$ 处连续，在其他所有点处都不连续。也就是说，它是一个只在一点连续的函数。1875 年，德国数学家卡尔·托马克（Karl Thomae）构造了一个更加怪异的函数：当 x 是有理数时，假设 x 的最简分数表达为 $\frac{n}{m}$，则令函数值为 $\frac{1}{m}$；当 x 是无理数时，令函数值为 0。这个函数的样子大概如图 1 所示，它有一个形象的别名——爆米花函数（popcorn function）[①]。爆米花函数拥有一个非常惊

① 一些数学书上也把它叫做“黎曼函数”，这是以德国数学家波恩哈德·黎曼（Bernhard Riemann）的名字命名的。

人的性质：它在所有有理点均不连续，在所有无理点均连续。这个例子告诉我们，在研究函数的连续性时，我们会遇到很多复杂的情形，千万不能凭直觉想当然地得出结论。更详细的描述可以在很多数学分析课本中找到。

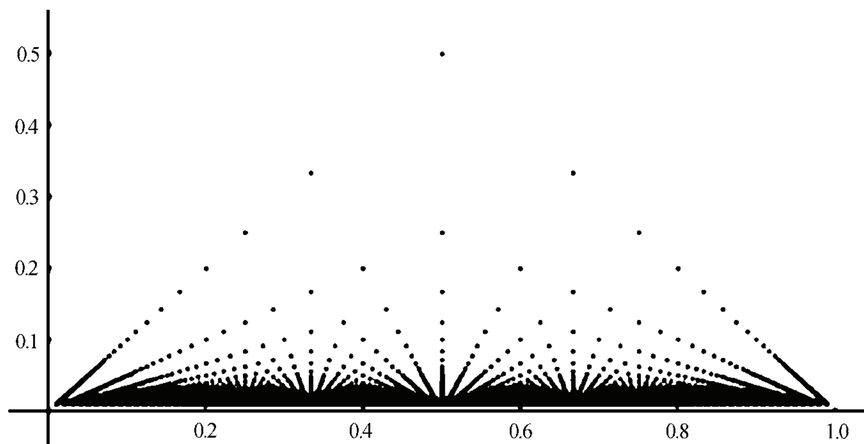


图 1

之后，我便开始收集满足各种奇异性质的函数：处处连续但处处不可导的函数，处处连续但只在一点可导的函数，连续单调递增但导数几乎处处为 0 的函数，连续单调递增并趋于某个上界但导数并不趋于 0 的函数，等等。但是，它们大多是有针对性的、精心构造的函数。我一直没能见到像狄利克雷函数那样简单而又霸气的构造。直到某日，我见识了数学家约翰·康威提出的康威十三进制函数。它是一个从全体实数集映射到全体实数集的函数，其中每个实数都被映射了无穷多次！

康威十三进制函数远不止这点本事。比方说，这个函数虽然处处不连续，但在任意区间 $[a,b]$ 里，函数值都将取遍 $f(a)$ 和 $f(b)$ 之间的所有数。再比方说，这个函数虽然处处有限，但在任意小的区间 $[a,b]$ 里，函数都是无界的。事实上，上面所有这些违反直觉的性质都来源于一个更强的、更不可思议的性质：在任意小的给定区间 $[a,b]$ 里，函数的值域都是整个实数域！这个函数的图像将会布满整个平面直角坐标系，在平面上任意选择一个任意小的区域，我们都能在里面找到该函数的一个点！

康威十三进制函数 $f(x)$ 是这样定义的。首先，把 x 转换为一个十三进制的无限小数（如果是有限小数，可以把它看做一个后面跟有无穷多个数字 0 的无限小数），然



后取出它的小数部分。这个小数部分应该是一个由 1、2、3、4、5、6、7、8、9、0、A、B、C 十三种符号组成的无限长的符号串，其中 A、B、C 三种符号是“非十进制符号”。如果这个符号串中非十进制符号的个数有限，并且最后一个非十进制符号是 C、倒数第二个非十进制符号是 A 或者 B，那么就去掉在这个 A 或者 B 以前的所有符号，然后把剩下的符号串视为一个十进制小数（把 A 当成正号，把 B 当成负号，把 C 当成小数点），作为 $f(x)$ 的函数值。其他情况下，$f(x)$ 一律为 0。

比方说，某个 x 是十三进制下的混循环小数 12A.AB3 C71 B67 C61 808 080…。由于它的小数部分有一个形如“符号 A 或者 B 加上一串十进制符号再加上一个 C 再加上一串无限长的十进制符号”的“后缀”（即 B67 C61 808 080…），那么我们把它提取出来，按规则把它理解成一个十进制小数，得到 −67.618 080 80… 。它就是 $f(x)$ 的值。

显然，对于任意一个给定的 y 以及任意小的一个区间 $[a,b]$，我们都能构造一个位于区间 $[a,b]$ 内的实数 x，使得它的十三进制小数展开中，从足够靠后的某个地方起，正好就是实数 y 的小数展开。这就证明了，在任意小的一段区间里，康威十三进制函数的值域都是全体实数。这个函数的图像遍布整个平面，形成一个在整个平面内稠密的点集。

在实分析中，大家会见到各种奇怪的函数，其中狄利克雷函数和康威十三进制函数恐怕是构造最简单、效果最拔群的函数了。不过，这趟“奇异函数”之旅并未结束。在下一节中，你将会看到一个更惊人的东西。

44. 塔珀自我指涉公式

你相信吗，一个不等式的图像里竟然写着这个不等式本身？2001 年，在介绍一种全新的方程图像绘制算法时，塔珀（Jeff Tupper）构造了这样一个有趣的不等式：

$$\frac{1}{2}<\left\lfloor \operatorname{mod}\left(\left\lfloor \frac{y}{17}\right\rfloor 2^{-17\lfloor x\rfloor-\operatorname{mod}(\lfloor y\rfloor,17)},2\right)\right\rfloor$$

其中，$\lfloor x\rfloor$ 表示向下取整，即不超过 x 的最大整数；$\operatorname{mod}(x, y)$ 则表示 x 除以 y 的余数。神奇的是，如果把满足不等式的点描绘在平面直角坐标系上，那么对于某个特殊的数 n，图像在 $0\leqslant x\leqslant 106$、$n\leqslant y\leqslant n+17$ 的范围内将会是图 1 所示这个模样。

图　1

这个 n 的值是：

4858450636189713423582095962494202044581400587983244549483093085061934704708809928450644769865524364849997247024915119110411605739177407856919754326571855442057210445735883681829823754139634338225199452191651284348332905131193199953502413758765239264874613394906870130562295813219481113685339535565290850023875092856892694555974281546386510730049106723058933586052544096664351265349363643957125565695936815184334857605266940161251266951421550539554519153785457525756590740540157929001765967965480064427829131488548259914721248506352686630476300[1]

① 杰夫 · 塔珀的论文原文中所给出的 n 值可能有误，这里给出的是正确的 n 值。



觉得这个很神奇吧？你也许会想，天哪，这个是怎么构造出来的啊！其实这一点都不神奇。塔珀自我指涉公式仿佛是数学里的魔术，当看穿了里面的把戏之后，你便能轻易构造出无数个这样的式子来，塔珀自我指涉公式也就没什么意思了。继续读下面的内容之前，大家不妨先思考一下。你能看出其中的奥秘吗？

在这个式子里，变量 x 和 y 出现的每个地方都加上了取整符号，因此整个图像都是一格一格的。于是，我们只需要考察 $\frac{1}{2}<\left\lfloor \operatorname{mod}\left(\left\lfloor \frac{y}{17} \right\rfloor 2^{-17x-\operatorname{mod}(y,17)},2\right)\right\rfloor$ 的整数解，把这些解描绘在平面直角坐标系上，再扩展成一幅像素画即可。另外，一个数乘以 2 的负 k 次方相当于对应的二进制数小数点左移 k 位（正如一个数乘以 10 的负 k 次方相当于这个十进制数的小数点左移 k 位），那么 $\left\lfloor \operatorname{mod}\left(N\cdot 2^{-k},2\right)\right\rfloor$ 实质上就是 N 的二进制数右起第 k 位上的数字（正如 $\left\lfloor \operatorname{mod}\left(N\cdot 10^{-k},10\right)\right\rfloor$ 可以提取出十进制数 N 右起第 k 位上的数字）。当 $x=0,1,2,3,\cdots$ 并且 $y=17N,17N+1,17N+2,\cdots,17N+16$ 时，指数 $-17x-\operatorname{mod}(y,17)$ 恰好对应 $0,-1,-2,\cdots,-17,-18,-19,\cdots,-34,-35,-36,\cdots$，于是位于 $y=17N$ 和 $y=17(N+1)$ 之间的图像的每个像素和 N 的二进制中的每一位数字一一对应。

随着 N 值的增加，图形的像素会一点一点地变化。当纵坐标足够大时，必然会出现一段高度为 17 的图像，图像的样子和不等式本身的样子完全相同。

当然，我们也可以把塔珀自我指涉公式中的 17 改成任何你想要的数。图 2 给出了 $\frac{1}{2}<\left\lfloor \operatorname{mod}\left(\left\lfloor \frac{y}{3} \right\rfloor 2^{-3\lfloor x\rfloor-\operatorname{mod}(\lfloor y\rfloor,3)},2\right)\right\rfloor$ 的图像，可以看到，随着纵坐标的增加，图像依次枚举了所有高度为 3 的黑白像素画。

因而，你可以在塔珀自我指涉公式中“找到”任何你想要的图像，只需要适当选取图像高度，把图像编码为二进制数，并转换为十进制数即可。你甚至可以告诉你的恋人，说你发现了一个函数，函数在某个位置的图像正好是“某某某我爱你”的字样！

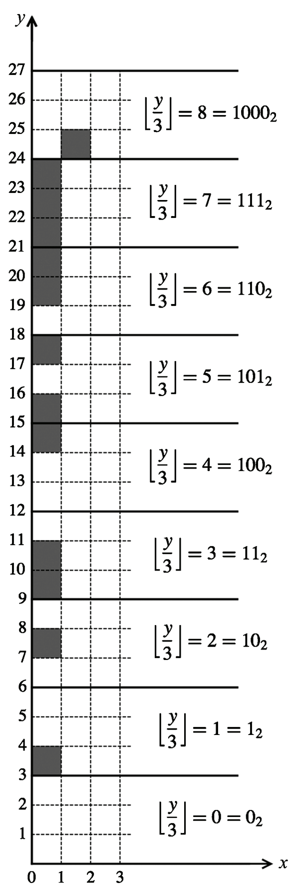

y
27
26
25
24
23
22
21
20
19
18
17
16
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
1
2
3
x
$\left\lfloor \frac{y}{3} \right\rfloor = 8 = 1000_2$
$\left\lfloor \frac{y}{3} \right\rfloor = 7 = 111_2$
$\left\lfloor \frac{y}{3} \right\rfloor = 6 = 110_2$
$\left\lfloor \frac{y}{3} \right\rfloor = 5 = 101_2$
$\left\lfloor \frac{y}{3} \right\rfloor = 4 = 100_2$
$\left\lfloor \frac{y}{3} \right\rfloor = 3 = 11_2$
$\left\lfloor \frac{y}{3} \right\rfloor = 2 = 10_2$
$\left\lfloor \frac{y}{3} \right\rfloor = 1 = 1_2$
$\left\lfloor \frac{y}{3} \right\rfloor = 0 = 0_2$

图　2



45. 俄罗斯方块可以永无止境地玩下去吗？

大家在玩俄罗斯方块的时候有没有想过这样一个问题：如果玩家足够厉害，是不是永远也不可能玩死？换句话说，假设你是万恶的游戏机，你打算害死你面前的玩家，你知道任意时刻游戏的状态，并可以有针对性地给出一些明显不合适的方块，尽量迫使玩家面对最坏的情况。那么，你有没有一种算法能保证害死玩家呢？或者，会不会玩家无论如何都存在一种必胜策略呢？注意，俄罗斯方块的游戏区域是一个宽为 10，高为 20 的矩形，并且玩家可以事先看到下一个给出的方块是什么。在设计策略时必须考虑这一点。

相信很多人有过这样的经历：玩俄罗斯方块时一开局就给你一个 S 形方块，让完美主义者感到异常别扭。结果，第二个方块还是 S，第三个方块依旧是 S，相当令人崩溃。于是，我们开始猜测，如果游戏机给你无穷个 S 形方块，玩家是不是就没有解了？答案是否定的。如图 1 所示，从第 10 步开始，整个局面产生一个循环；只要机器给的一直都是 S 形方块，玩家可以不断重复这几个步骤，保证永远也死不了。

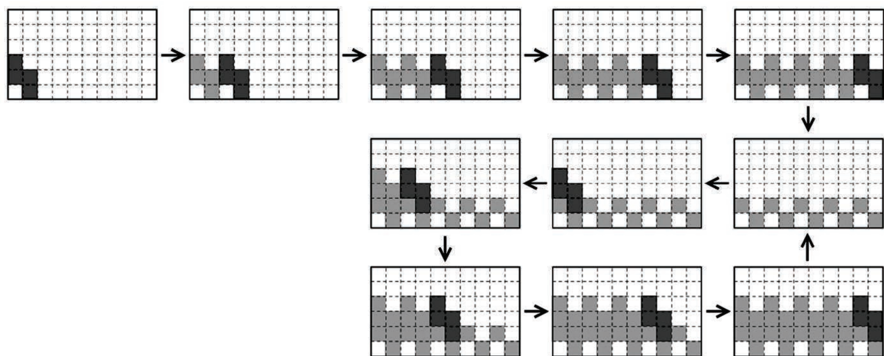


图 1

不过，这个循环是在游戏场地清空了的情况下才产生的。有人会进一步想了，要是在玩着玩着，看着你局势不好时突然给你无穷多个 S 形方块呢？事实上，此时局面的循环依然可能存在，如图 2 所示。在第 5 个 S 形方块落地后，循环再次产生。

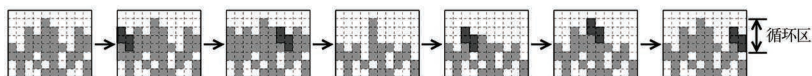


图　2

俄罗斯方块究竟是否存在必死的情况呢？1988 年，约翰·布茹斯托斯基（John Brzustowski）的一篇论文给出了肯定的答案。他给出了一种算法，可以保证游戏机能够害死玩家，即使要求它必须提前向玩家展示下一个方块的形状。构造的关键在于，整个游戏的局面个数是有限的（2 的 200 次方），如果玩家一直不死，在某一时刻必然会重复某一状态。我们把两次重复状态及其之间的游戏过程叫做一个“循环”，这个循环实际影响到的那些行就叫做“实际循环区”。例如，图 2 就是一个循环，这个循环的“实际循环区”是从第 4 行到第 7 行这四行。

我们把宽为 10 的游戏区域划分为 5 个宽为 2 的“通道”，从左至右用 1 到 5 标号。注意到图 1 和图 2 中的两个循环都有一个共同点：每个 S 形方块最终都完全落在某个通道内。事实上，对于任意一个只有 S 形方块的循环，我们都能得出这个结论。也就是说，如果游戏机一直给你 S 形的方块，你却用它们弄出了一个循环，那只有一种可能：所有 S 形方块的下落位置都没有跨越通道（就像图 3 中的方块 A、B 那样，而非方块 C、D 那样）。

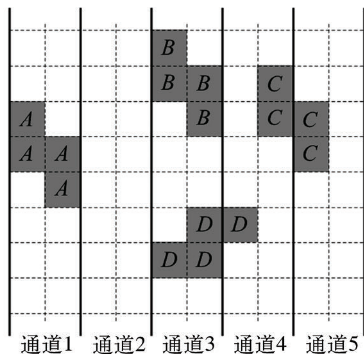


图　3



为了证明这一点，我们对通道编号实施归纳。考虑命题 $P(x)$：如果某个 S 形方块（或它的一部分）落在了通道 x 的左边（或者说前 $2(x-1)$ 列里），那它一定完全落在某个通道内。$P(1)$ 显然成立：方块根本不可能占据通道 1 左边的某个格子，因为通道 1 左边什么都没有。下面我们说明，当 $P(n)$ 为真时，$P(n+1)$ 也为真。

我们首先要证明一个引理：在循环中的任意时刻，通道 n 的实际循环区内绝对不可能出现形如“□■”的两个并排的格子。如图 4(1)所示，假设图中星号方块所在行是通道 n 的实际循环区内位置最低的“□■”的结构。假如这一行被消掉了，又由归纳假设，不存在哪个 S 形方块跨越了该通道的左边界，因此只有一种可能：某个 S 形方块从左侧面挤了进来，如图 4(2)所示。但这样一来，我们又产生了一个更低的“□■”，矛盾。这就是说，星号方块所在行一直没被消去。但这也是不可能的，因为实际循环区内是一个新陈代谢、以旧换新的更替过程，每一行最后都是会被消除的。

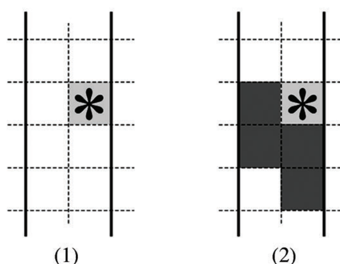


图 4

接下来，考虑命题 $P(n+1)$。要想让 S 形方块占据通道 n 的格子，只有图 5 这四种情况。但是，由于我们之前证明了通道 n 中不能存在“□■”，因此在这个 S 形方块落下之前，星号方块都是已经存在的了。注意到，每一个 S 形方块的下落都致使“■□”形结构减少，但第一种情形除外——它消除了一个“■□”形结构，但给其上方带来了一个新的，所以“■□”形结构个数保持不变。没有哪种情形能够增加“■□”的个数。但是，通道 n 的“■□”形结构个数应该是恒定的，因为它在一个循环区里。因此，只有第一种情况才能够被接受。

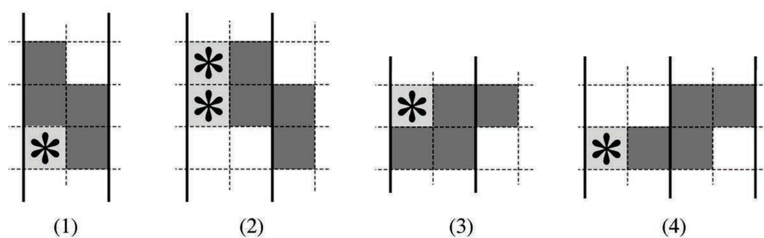


图 5

也就是说,仅含有 S 形方块的循环只有一种情况——S 形方块在各个通道内重叠,填满并消除若干行后回到初始状态。实际循环区内的每个通道都是一个模样:底下是 0 个或多个“■■”,顶部一个“■□”。注意,最右侧那个通道的最顶端是一个“■□”,右边这个空白永远也不可能用 Z 形方块填上。也就是说,在一个只含 S 形方块的循环区内,必然会有某一行,它的最右侧是一个“■□”,它保证了该行不能仅用 Z 形方块消掉。如图 6 所示,箭头所指的行无法单用 Z 形方块消除,因为星号位置不可能用 Z 形方块填充。

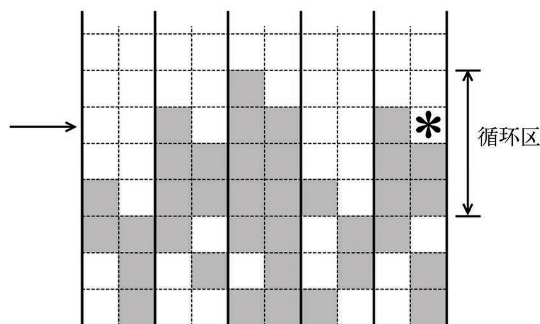


图 6

下面我们给出游戏机害死玩家的算法。

(1) 不断给出 S 形方块并显示下一个方块也为 S 形,直到出现一个循环。

(2) 给出一个 S 形方块并显示下一个方块为 Z 形。

(3) 不断给出 Z 形方块并显示下一个方块也为 Z 形,直到出现一个循环。

(4) 给出一个 Z 形方块并显示下一个方块为 S 形。

(5) 跳回(1)并重复执行。



这样的话，玩家为什么会无解呢？由上面的结论，在第(1)步后，游戏区域中出现了一个不能用 Z 形消除的行。即使再给你一个 S 形方块，这一点仍然无法挽救，因为填充星号空格的唯一途径就是插一个 S 形进去，但这立即又产生了一个 Z 形永远放不进去的空位（如图 7 所示）。然后，玩家就拿到了一大堆 Z 形，最终必然会产生另一个循环区，且这个循环区在刚才那个无法消去的行上（循环区不可能包含一个不能消除的行，因为正如前面所说，一个实际循环区的所有行最终都是会被消掉的，这样才可能循环）。这个循环区的最左边那个通道将会产生一个“□■”结构，是 S 形方块所不能消去的。于是，游戏机又给出一大堆的 S，最终使得两种无法消去的行交替出现，直至游戏结束。

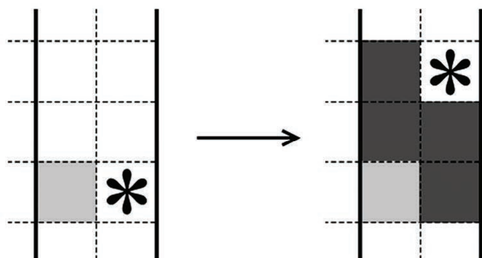


图 7

到此为止，我们便完成了整个证明。有人或许会指出：现实的游戏机并没有主观能动性啊？事实上，即使方块是随机出的，如果你倒霉到家了，这种特殊的方块序列可能恰好就让你一个不错地碰上了。虽然这种怪事的发生概率非常非常低，但理论上说毕竟是有可能的，因此俄罗斯方块终究不是玩不死的，总有一个时刻会 Game Over。

有趣的是，这个结论还可以直接扩展到场地为任意宽度的俄罗斯方块游戏。当场地宽为其他偶数时，上述证明同样有效；当场地宽为奇数时，无穷多个方形方块就可以直接干掉玩家。



46. 无以言表的大数：古德斯坦数列

我们刚才见识了很多大数。不过，比起古德斯坦（Goodstein）数列来，就算小巫见大巫了。

一个 b 进制的数总可以写成 $a_nb^n+a_{n-1}b^{n-1}+\cdots+a_1b+a_0$ 的形式，其中 $a_0,a_1,\cdots,a_n$ 都是小于 b 的数。四进制数 102 312 可以写成 $1\times4^5+2\times4^3+3\times4^2+1\times4+2$，正如十进制数 1206 可以写成 $1\times10^3+2\times10^2+6$ 一样。麻烦的是，在 102 312 的四进制数展开式中，有 1×4^5 这么一项，而指数 5 却并不是一个四进制展开式。对于一个完美主义者来说，或许要把 1×4^5 里的那个 5 也改写成 $4+1$，这才算是“真正的”四进制展开。同理，二进制数 1 000 011 写成 2^6+2+1 还不够，$2^{2^2+2}+2+1$ 才是真正的二进制展开。当然，倘若指数的指数还有太大的数字，我们应该继续对其进行展开，直到整个展开式只由不超过底数 b 的数字组成。不妨把这种 b 进制的展开方式叫做“完全 b 进制展开”。

给定一个初始的正整数 m，我们可以根据下面的规则生成古德斯坦数列 $G_i(m)$。其中，$G_1(m)$ 就等于 m 本身，$G_2(m)$ 则是把 $G_1(m)$ 的完全二进制展开中所有的数字 2 都替换成 3 后再减去 1 所得的数，$G_3(m)$ 则是把 $G_2(m)$ 的完全三进制展开中所有的数字 3 都替换成 4 后再减去 1 所得的数，以此类推。比方说，当 $m=8$ 时：

$$G_1(8)=8=2^{2+1}$$

$$G_2(8)=3^{3+1}-1=80=2\times3^3+2\times3^2+2\times3+2$$

$$G_3(8)=2\times4^4+2\times4^2+2\times4+2-1=533=\cdots$$

……

古德斯坦定理指出，不管初始数 m 是多少，按照上述方法迭代下去，最后总有一个时刻会变成 0。例如，当 $m=3$ 时：

$$G_1(3)=3=2+1$$



$G_2(3)=3+1-1=3$

$G_3(3)=4-1=3$

$G_4(3)=3-1=2$

$G_5(3)=2-1=1$

$G_6(3)=1-1=0$

6 步之后，数列便收敛到了 0。不过，这并不稀奇。到了第 3 步，展开式中已经没有可以替换的数了；在此之后，数列开始递减，很快便成了 0。

但是，当 $m=4$ 时，情况就大不一样了：

$G_1(4)=4=2^2$

$G_2(4)=3^3-1=26=2\times3^2+2\times3+2$

$G_3(4)=2\times4^2+2\times4+2-1=41=2\times4^2+2\times4+1$

$G_4(4)=2\times5^2+2\times5+1-1=60=2\times5^2+2\times5$

$G_5(4)=2\times6^2+2\times6-1=83=2\times6^2+6+5$

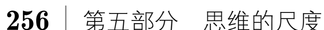
……

这个数列将会持续增长到几百，然后增长到几千，然后增长到几万、几亿，然后增长到几百位数、几千位数甚至上亿位数。但根据古德斯坦定理，数列最终总会回到 0 的，只不过花费的时间有可能会相当长。事实上，一直要到第 $3\times2^{402\ 653\ 211}-2$ 步，数列才会变成 0。$3\times2^{402\ 653\ 211}-2$ 步！这是一个拥有上亿位数字的超级大数！

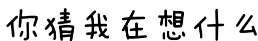
我们通常用 $\mathcal{G}(m)$ 来表示初始值为 m 时迭代到 0 所需要的步数。我们已经知道了 $\mathcal{G}(3)=6$，而 $\mathcal{G}(4)=3\times2^{402\ 653\ 211}-2$，可见函数 $\mathcal{G}(m)$ 增长速度之快。$\mathcal{G}(5)$ 则是一个更大的数，它的位数将会远远超过 $\mathcal{G}(4)$。（注意，并不是说它的位数超过了 $\mathcal{G}(4)$ 的位数，而是它的位数超过了 $\mathcal{G}(4)$。）

那么 $\mathcal{G}(8)$ 呢？这将会是一个非常非常巨大的数，即使说它有多少位，或者它的位数有多少位，或者需要在前面那句话里嵌套进多少个“的位数”，也无法表达出它的大小来。

除非……除非我们自创一种大数表示法。



47. 乘法之后是乘方，乘方之后是什么？



《哥德尔、艾舍尔、巴赫：集异璧之大成》一书的作者曾经说过，他小时候曾经有一个最让他激动的想法：3+3+3，用3个3和自身运算！

但我小时候并不觉得这很令人激动，因为我很早就知道，3个3相加就是3乘以3。但我好奇的是，这会有什么样的应用题呢？每次买3个苹果，连续买了3次，问一共买了多少个苹果？这听上去似乎不太自然。

后来我才知道，长方形的面积计算就是乘法应用最常见的例子。而倍数关系的表达更是让我大开眼界——在比较两个相差甚远的数量时，我们可以利用乘法的关系，



直接使用“我比你多多少倍”的句型！

于是我自然往下想了下去，如果拿 3 个 3 和自己相乘，会得到什么呢？后来我知道了乘方的概念。乘方，或者叫幂，这个概念并不是自古就有的。古希腊人发明了平方和立方，但只用于面积、体积的计算。在当时，4 次方、5 次方是没有任何实际意义的。随着人类文明的进步，人们需要应付的数字也越来越庞大。重复对折纸张、增长率的叠加和赌博游戏中的翻番都会涉及相同数量的连乘。于是，到了文艺复兴时期，数学家们开始用乘方来表示把同一个数连乘多次的结果，a^b 就表示 b 个 a 相乘。和科技、建筑、能源、生产力一样，发明大数记号也成了人类历史中不可缺少的一环。

利用乘方的记号，我们已经能表示宇宙中几乎所有有意义的数了，整个宇宙的基本粒子数量也不过10^{80} 。

但不知大家是否曾想过，乘方之上究竟是什么？

很容易想到，比乘方更大一级的运算就是把 b 个“a 次方”重叠起来。不过，这里我们却遇到了一个之前不曾遇到的问题：a^{a^a} 究竟应该等于$\left(a^a\right)^a$，还是$a^{\left(a^a\right)}$？我们不妨亲自算一算，不同算法得到的结果相差有多远：

$$\left(2^2\right)^2=4^2=16$$

$$2^{\left(2^2\right)}=2^4=16$$

难道用两种不同的计算顺序得到的结果总是相同的吗？换 $a=3$ 试试：

$$\left(3^3\right)^3=27^3=19\,683$$

$$3^{\left(3^3\right)}=3^{27}=7\,625\,597\,484\,987$$

哇，这下可就差远了。可以想象，如果把“a 次方”再多迭代几次，从右往左算和从左往右算会差得更多。恐怖的是，我们通常约定，当有多重指数时，运算正是按照从右往左算的顺序进行的。试想，若有一种运算专门用来表示 b 个 a 构成的指数塔，这种运算的威力会有多大。

1947 年，当英国数学家鲁本 · 古德斯坦（Reaben Goodstein）研究前一节提到的那个序列时，他遇到了一些连乘方也无法表达出来的大数。于是，古德斯坦便正式提出了这种超越乘方的运算。他把 b 个指数 a 迭代的结果记为 ${}^ba$ ，也就是把 b 放在 a 的左上角（见图 1）。这也就是我们现在所说的“超级幂”。在国外的一些论坛上，有时

也能看见 a^^b 的表示方法，便于在纯文本格式下传播。

不过，当时古德斯坦并没有用超级幂(superexponentiation)一词，而是用的 tetration 一词。这是由前缀“四”（ tetra- ）和迭代（ iteration ）一词合成的，意即排在加法、乘法、乘方之后的第四级运算。事实上，tetration 比 superexponentiation 更常用一些。网上甚至有一个 tetration 论坛，论坛里活跃着一群热爱 tetration 的数学玩家。

$$^{b}a = a^{a^{\cdot^{\cdot^{a}}}} \Big\} b\text{个}a$$

图 1

超级幂是一个极为厉害的运算，它的增长速度非常惊人。在很小的数之间进行超级幂运算，就有可能得到一个巨大的天文数字。$^{3}2$ 等于 $2^{2^{2}}=16$ ，而 $^{4}2$ 就等于 $2^{2^{2^{2}}}=65\,536$。那么，$^{5}2$ 等于多少呢？它应当等于 2 的 65 536 次方，其结果是一个上万位的数。那 $^{6}2$ 呢？$^{100}100$ 呢？大家自己去想象吧。

这时，我们仿佛重新遇到了我小学时代的困惑：超级幂有什么用途？我们能用超级幂编出什么应用题来？刚才说到，古希腊人生活太简单，不知道乘方有什么实际意义。现在，我们自己似乎也变成了窘迫的古希腊人。是否随着人类文明的进一步发展，未来人会随手使用超级幂，并在某本数学书上分析 21 世纪的人类为什么还要如此吃力地发明超级幂呢？我觉得有可能，不过这并不重要。现在的我们已经认识到，数学发展的动力并不是解释生活中的现象，数学发展的动力是数学这个学科本身。超级幂在生活中没有实际意义，并不妨碍我们发明超级幂这个记号。

人类的想象力是无止境的。即使超级幂已经大到没有任何实际意义的地步，大家还是会问，再把“a 次超级幂”迭代 b 层（注意运算顺序仍是从最深那一层开始），又会得到什么？是否就得到了第五级的运算呢？或许你马上就意识到了，这样扩展上去是没有尽头的，每一级运算迭代之后都能产生更高一级的运算。虽然此时脑子已经有点乱了，但是数学语言的严格性和理想性告诉我们，利用某种清晰的数学符号和递归法则，我们一定有办法定义出等级越来越高的运算来。



$$\left.{}^{a^{\cdot^{\cdot^{a}}}}a\right\}b\text{个}a = ?$$

图 2

古德斯坦厉害就厉害在这儿。他定义了古德斯坦记号 $G(n,a,b)$ ，以此表示 a 与 b 之间的第 n 级运算。当 $n=0$ 时，规定 $G(0,a,b)=b+1$ 。也就是说，第 0 级运算是一个一元运算——自然数的后继。当 $n=1$ 时，规定边界值 $G(1,a,0)=a$ ，并规定 $G(1,a,b)$ 表示对 $G(1,a,0)$ 的值进行上一级操作（后继操作），并重复迭代 b 次，其结果也就是 a 加上 b 。一般地，有：

$$G(n,a,b)=G(n-1,a,G(n,a,b-1))$$

其中边界值为：

$$G(1,a,0)=a$$

$$G(2,a,0)=0$$

$$G(3,a,0)=1$$

$$G(4,a,0)=1$$

$$G(5,a,0)=1$$

……

这就形式化地给出了第 n 级运算的意思。

其实，类似的东西不止一次地被提出过。高德纳（Knuth）箭头记号也是一种常用的大数表示方法，其思想与古德斯坦记号几乎完全一样。阿克曼（Ackermann）函数也是一个神速增长的函数，它的定义也有异曲同工之处。很多外文数学论坛则用 $a[n]b$ 来表示 a 与 b 之间的第 n 级运算，是我比较喜欢的一种符号。

当然，有 $a[n]b$ ，必然会有 $a[a[n]b]b$ ，从而又会有 $a[a[a[n]b]b]b$ ……没有最大的数，只有更大的数。人脑和数学是两个神奇的东西，没有什么数大到人脑想不出来，也没有什么数大到数学表示不出来。仅仅在脑中试想一下 $100[100]100$ ，你的思维就已经超越了整个宇宙的大小了。

48. 不同维度的对话：带你进入四维世界

人的思维能够超越宇宙的大小，这并不奇怪。事实上，人的思维还能超越宇宙的维度。借助类比的思想，我们可以在大脑中勾勒出一幅四维世界的景象。

生活在三维世界的我们，确实很难理解四维空间。正如我们很难告诉二维世界的人，三维空间是什么样子的。

现在，假设我是一个二维世界的人，我不能理解什么是“高度”，什么是“体”，什么是“空间”。你想向我描述三维世界中的立方体。你该怎么说呢？你或许会从立方体的展开图开始谈起：图 1 就是一个立方体的展开图，如果我们剪一个这种形状的纸板，就可以把它折成一个正方体。我不理解了。

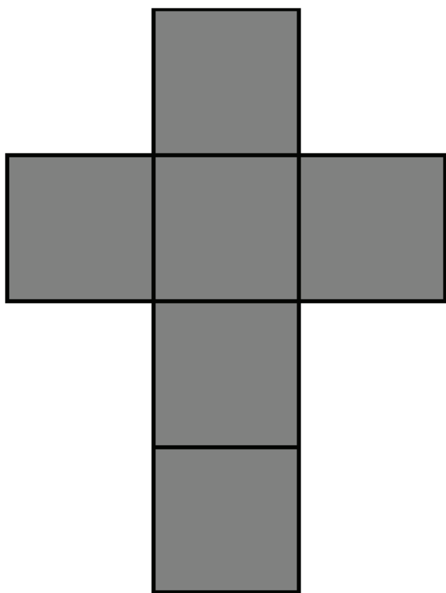


图　1



你说说该怎么做呢？

先把上面几个正方形折起来，把对应的边粘在一起……

等会儿呢等会儿呢，这几个正方形是稳定的形状呀，它们的边怎么可能挨到一起呢？

傻了吧！在二维世界中它们不是活动的，但是它们可以向第三维度弯折啊！画个图 2 给你看吧，这就是把上面那几个正方形粘合起来的样子，这就成了一个没有封顶、还差一面的正方体……

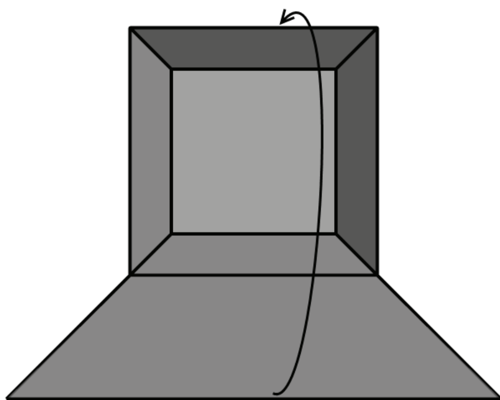


图 2

你耍赖！你这样弯折了之后正方形就不是正方形了，都变成梯形了！

不对，它们仍然是正方形。图 2 的 6 块区域其实都是正方形，只是由于透视作用，它们看上去好像变“斜”了。

嗯，好吧，你继续。

现在我们得到的是一个有盖的盒子。上面 5 个正方形（其中有 4 个由于处于第三维度而变了形）的“内部”已经形成“空间”了，可以往里面放东西了。要想做成一个封闭的正方体，只需要把剩下的那个正方形合上去就行了，最终结果就像图 3 那样。

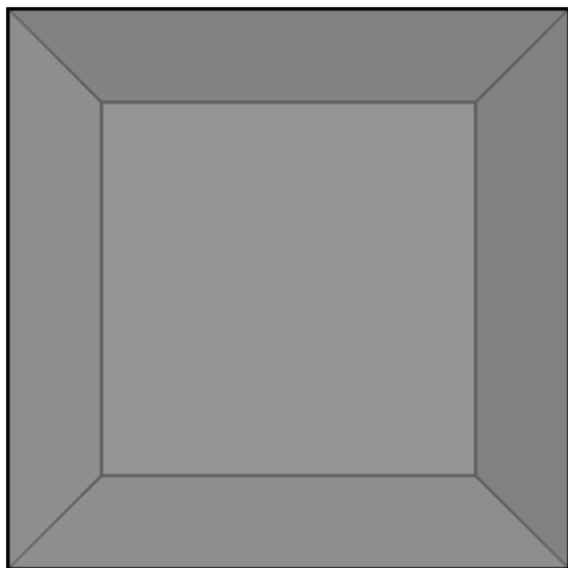


图　3

咦？图 3 里面，刚才最后要合上去的那个正方形到哪儿去了？

它就是最大的那个正方形。

胡说！那个大正方形是 5 个小正方形拼成的！这个大正方形刚才在图 2 里也有！

不是的。图 2 里的大正方形的确是 5 个小正方形拼成的轮廓，但图 3 里的那个大正方形是真实存在的，它就是最后合上去的那一块。这个大正方形也并不是和那 5 个小正方形重叠在一起，它们在第三维中的层次是不同的。图 3 就是你梦想的那个正方体了，它由 6 个正方形组成。你在图 3 中看到的一个小正方形，一个大正方形，四个梯形事实上都是正方形，而且它们都一样大。这 6 个正方形围成了中间的那个“空间”。

我还是不明白。那个大正方形也是在第三维度的，为什么它没变形呢？

这是因为，这个正方形是正对着我们的，它所在的方向不是第三个维度，因此看上去和原来一样。

那同一个方向上为什么又有一大一小两个正方形呢？



唉，真麻烦。这是因为，它们的朝向虽然一样，但在第三维度上的位置不一样。小的那个正方形在第三个维度离我们远一些，看起来就要小一些。

哦！我有点明白了。是不是说，旁边一圈那 4 个“正方形”是跨越了第三维的，因此在第三维空间中一部分离我们近，一部分离我们远，于是看上去就是由大到小渐变过去的，就像是变形了。

对！你理解得很好！说真的，平时生活在三维空间中，我都还没仔细想过这一点呢。

我好像真的明白了，说错了不要笑我哦。那个“空间”啊，说穿了就是大正方形擦着 4 个变形正方形在第三维度上向远处的小正方形移动所产生的“轨迹”。

正是正是！

哎呀我彻底明白了。怪不得我们说 n 维立方体有 2^n 个顶点呢，其实道理很简单。只需要把 $n-1$ 维立方体复制一份，然后把对应的顶点相连就可以了。这就是 $n-1$ 维立方体在第 n 维发生位移的结果，新增的那 2^{n-1} 条边就是点的轨迹。

太棒了！就是这样！我还给你看一个好玩的东西，让你看看三维立方体是如何旋转的。如图 4 所示，睁大眼睛仔细看好每个正方形都变到哪儿去了。

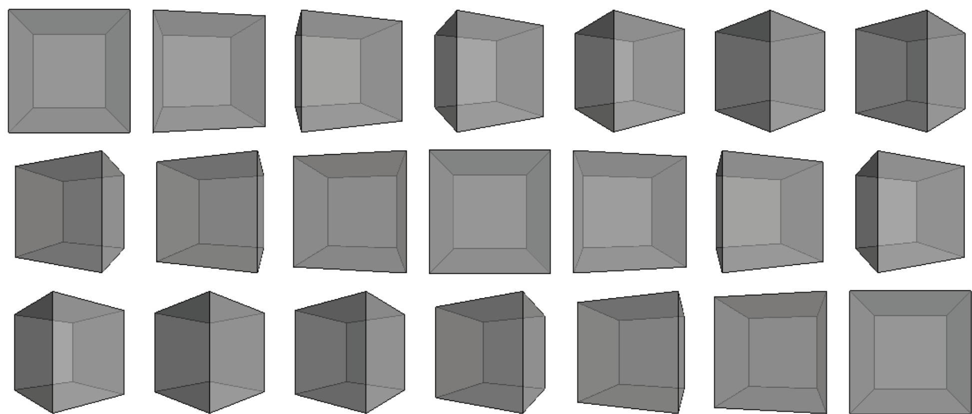


图 4

我又糊涂了。为什么从第二幅图变成第三幅图时，远处的小正方形能够穿越左边界，让其中一小半跑到边界左边来?

这个确实不好理解。小正方形并没有“穿过”那条竖直的边，那条边在第三维上离我们更近，而它在我们这个方向上的投影又与小正方形重合了。其实你可以看到，它们之间的拓扑关系仍然是不变的。

哦，于是乎远处的小正方形就转到侧面去了，然后又转到离我们近的位置来了，替代了原先大正方形的位置……

回去没事多想想吧。期待你睡觉时能够做出一个三维的梦。

好的。谢谢你让我懂得了三维空间。看来，二维世界的人理解三维空间真不容易啊!

好了，回到现实中来。这次，让我们交换一下位置，由我来描绘一个四维立方体的样子吧。你会发现，现在，一切都比你想象中的更容易了。

四维立方体是由 8 个大小相同的三维立方体组成的，其展开图如图 5 所示。

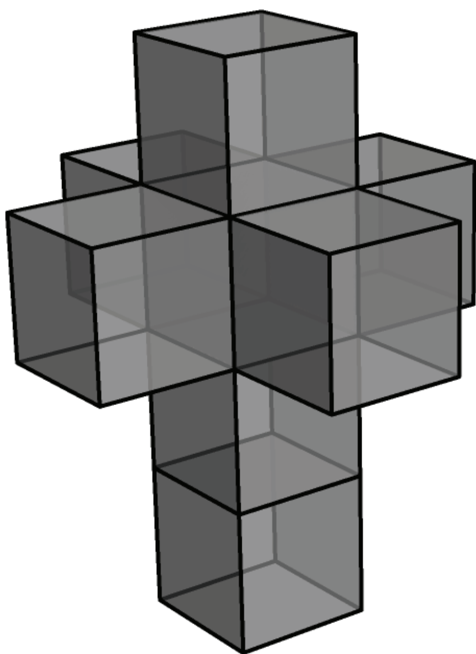


图　5

图 6 是粘合出来的四维盒子，还差一个盖子没有盖。这些看起来像棱台的东西其实都是根正苗红的正方体，只是由于它们在四维空间中位置不同，发生了透视。

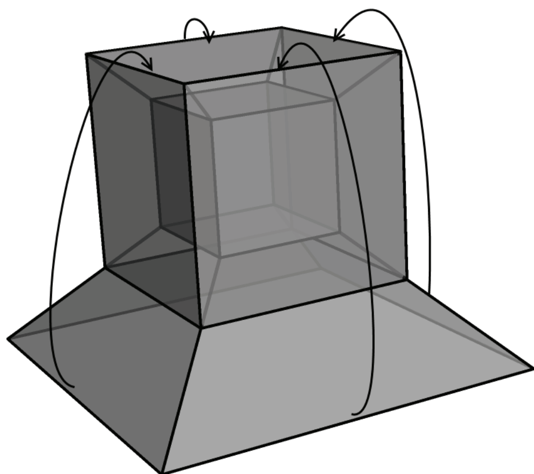


图　6

把盖子盖上后，我们就看到了传说中的四维立方体，如图 7 所示。

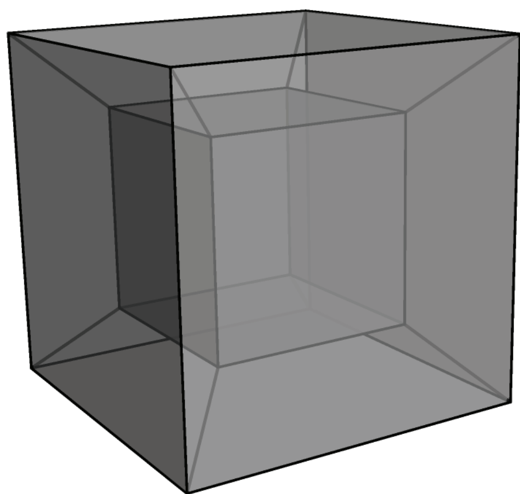


图　7

相信不少人都已经在其他地方见过这个图形了。图上有一大一小两个标准模样的立方体，这是第四维度上位置不同但都正对我们的两个“三维面”。其他棱台其实都

是正方体，只是看上去因透视而变形。四维立方体可以看做三维立方体的移动轨迹，因此画一个四维立方体很简单：画两个三维立方体，然后连接对应顶点即可。如图 8 所示，观察四维立方体的旋转，你会看到里面的小立方体穿过一个面跑到了外面，接下来还将继续变成最外面的大立方体。这一切都和二维向三维的推广是类似的。仔细观察思考，你还会发现更多可以类比的地方。

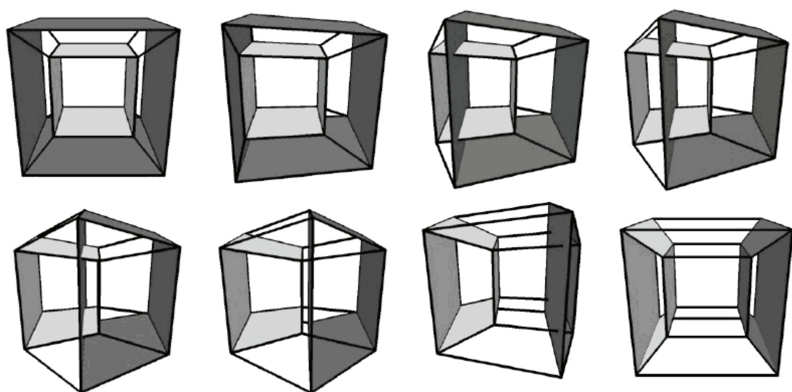


图　8

现在，合上书，闭上眼，体会一下超越三维空间的美妙感吧。

祝愿你今晚能够做一个四维的梦。

欢迎加入

图灵社区

电子书发售平台

电子出版的时代已经来临，在许多出版界同行还在犹豫彷徨的时候，图灵社区已经采取实际行动拥抱这个出版业巨变。相比纸质书，电子书具有许多明显的优势。它不仅发布快，更新容易，而且尽可能采用了彩色图片（即使有的书纸质版是黑白印刷的）。读者还可以方便地进行搜索、剪贴、复制和打印。

图灵社区进一步把传统出版流程与电子出版业务紧密结合，目前已实现作译者网上交稿、编辑网上审稿、按章发布的电子出版模式。这种新的出版模式，我们称之为“敏捷出版”，它可以让读者以较快的速度了解到国外最新技术图书的内容，弥补以往翻译版技术书“出版即过时”的缺憾。同时，敏捷出版使得作、译、编、读的交流更为方便，可以提前消灭书稿中的错误，最大程度地保证图书出版的质量。

开放出版平台

图灵社区向读者开放在线写作功能，协助你实现自出版的梦想。你可以联合二三好友共同创作一部技术参考书，以免费或收费的形式提供给读者，这极大地降低了出版的门槛。成熟的书稿，有机会入选出版计划，同时出版纸质书。

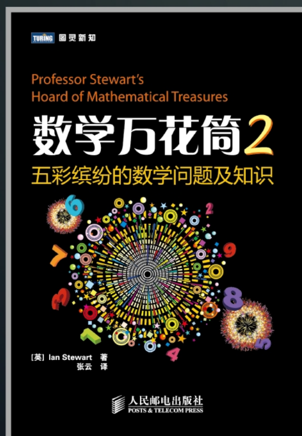
图灵社区引进出版的外文图书，都将在立项后马上在社区公布。如果有意翻译哪本图书，欢迎来社区申请。只要通过试译的考验，即可签约成为图灵的译者。当然，要想成功地完成一本书的翻译工作，是需要有坚强的毅力的。

读者交流平台

在图灵社区，读者可以十分方便地写作文章、提交勘误、发表评论，以各种方式与作译者、编辑人员和其他读者进行交流互动。提交勘误还能够获赠社区银子。欢迎大家积极参与社区开展的访谈、审读、评选等多种活动，赢取银子，可以换书哦！


TURING 图灵新知
双色印刷
如何破解达芬奇密码?
35问揭示数学之美
原名：How to Build a Brain
And 34 Other Really Interesting
Uses of Mathematics
[英] Richard Elwes 著
刘熙 译
人民邮电出版社
POSTS & TELECOM PRESS


TURING 图灵新知
1930
1742
1200
825
20个大问题
影响数学发展的
[英] Tony Crilly 著
王耀杨 译
THE BIG QUESTIONS: MATHEMATICS
人民邮电出版社
POSTS & TELECOM PRESS


TURING 图灵新知
Professor Stewart's
Hoard of Mathematical Treasures
数学万花筒2
五彩缤纷的数学问题及知识
[英] Ian Stewart 著
张云 译
人民邮电出版社
POSTS & TELECOM PRESS


TURING 图灵新知
你不可不知的
50个数学知识
50 Mathematical Ideas You Really Need to Know
[英] Tony Crilly 著 王悦 译
人民邮电出版社
POSTS & TELECOM PRESS

思考的乐趣

Matrix67数学笔记

“本书一大特色，是力图把道理说明白。作者总是用自己的语言来阐述数学结论产生的来龙去脉，在关键之处还不忘给出饱含激情的特别提醒。数学的美与数学的严谨是分不开的。数学的真趣在于思考……本书讲了不少相当深刻的数学工作，其推理过程有时曲折迂回，作者总是不畏艰难，一板一眼地力图说清楚，认真实践着古人“诲人不倦”的遗训。这个特点使本书能够成为不少读者案头床边的常备读物，有空看看，常能有新的思考，有更深的理解和收获。”

——张景中，中国科学院院士

“事实上顾森的每篇文章都在向读者展示数学确实好玩。数学好玩这个命题不仅对懂得数学奥妙的数学大师成立，对于广大数学爱好者同样成立。”

——汤涛，《数学文化》期刊联合主编，香港浸会大学数学讲座教授

图灵社区：www.ituring.com.cn
新浪微博：@图灵教育　@图灵社区
反馈/投稿/推荐信箱：contact@turingbook.com
热线：(010)51095186转604

分类建议　数学/科普读物

人民邮电出版社网址：www.ptpress.com.cn



ISBN 978-7-115-27586-8
定价：45.00元

看完了

如果您对本书内容有疑问，可发邮件至 contact@turingbook.com，会有编辑或作译者协助答疑。也可访问图灵社区，参与本书讨论。

如果是有关电子书的建议或问题，请联系专用客服邮箱：
ebook@turingbook.com。

在这可以找到我们：

微博 @图灵教育：好书、活动每日播报
微博 @图灵社区：电子书和好文章的消息
微博 @图灵新知：图灵教育的科普小组
微信 图灵访谈：ituring_interview，讲述码农精彩人生
微信 图灵教育：turingbooks